Complications in Causal Inference: Incorporating Information Observed After Treatment is Assigned

Complications in Causal Inference: Incorporating Information Observed After Treatment is Assigned

A dissertation presented

by

David Allan Watson

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May 2014

Dissertation Advisors: Professors Donald B. Rubin and Joseph K. Blitzstein David Allan Watson

Complications in Causal Inference: Incorporating Information Observed After Treatment is Assigned

Abstract

Randomized experiments are the gold standard for inferring causal effects of treatments. However, complications often arise in randomized experiments when trying to incorporate additional information that is observed after the treatment has been randomly assigned. The principal stratification framework has provided clarity to these problems by explicitly considering the potential outcomes of all information that is observed after treatment is randomly assigned. Principal stratification is a powerful general framework, but it is best understood in the context of specific applied problems (e.g., non-compliance in experiments and "censoring due to death" in clinical trials). This thesis considers three examples of the principal stratification framework, each focusing on different aspects of statistics and causal inference.

In particular, the first example considers early escape designs in which additional rescue medication is provided for patients that do not respond well to the assigned treatment in a placebo-controlled clinical trial. We demonstrate complications that arise in such trials as well as provide a Bayesian analysis of a dataset with such complications. Another example considers the case of binary outcomes in a randomized experiment. Binary outcomes, in combination with a binary treatment, necessarily lead to four principal strata that cannot be identified with the observed data. We con-

sider inference for the average causal effect by testing null hypotheses that determine the number of units in the principal strata of interest. Fisher's randomization test, a standard randomization-based analysis, breaks down for such hypotheses because they are not sharp and rely on nuisance unknowns. We interpret the randomization test as a Bayesian posterior predictive check, which can integrate out the nuisance unknowns. The last example focuses on estimands that assess the efficacy of a prophylactic treatment of HIV, which fits into the more general framework of assessing causal effects of treatment for preventing infectious diseases. We focus on two issues involving information observed after treatment is assigned: exposure to the disease and interference between units. We link these two issues by showing how interference occurs because an effective treatment reduces the exposure in the population.

Contents

Acknowledgments

I would like to acknowledge my advisors, Donald Rubin and Joseph Blitzstein, who have provided endless inspiration for statistical research and pedagogy. I would like to thank Luke Miratrix for reading my dissertation. I additionally thank the faculty, who have shaped my approach to statistical thinking.

I also want to thank the my fellow graduate students, who have offered hours of stimulating conversation and support. I especially thank Viviana García-Horton for her contributions to Chapter 1 of this dissertation.

I also thank Stacy Lindborg of Biogen Idec, who provided the inspiration and data for Chapter 1 of this work.

Lastly, I thank my family, who have always encouraged me to pursue my passions.

To Melanie Yuen.

Chapter 1

A principal stratification approach to receipt of rescue medication

1.1 Introduction

The ethics of placebo-controlled clinical trials versus active comparator studies is a long debated and controversial subject (Lasagna, 1979; Hill, 1994; Temple and Ellenberg, 2000). Early escape designs for placebo-controlled trials attempt to reduce the amount of time a subject takes the placebo in order to limit the inherent risk to less healthy patients of taking an inactive treatment (Temple, 1994). The design has a clearly defined protocol that allows subjects to receive an additional (non-trial) treatment, known as rescue medication, if they deteriorate with their assigned treatment. Aspects of early escape designs are commonly employed in drug development trials. The Food and Drug Administration (FDA) recognizes early escape designs, but warns that such designs may only provide information on short-term effective-

ness (FDA, 2001, Section 2.1.5.2.2). As Temple (1994) noted, early escape designs are "truly randomized for a particular drug-placebo comparison only [up to the first time rescue medication is available], after which analysis is complicated by outcome-dependent changes in therapy".

To get a sense of these complications, consider the standard intention-to-treat (ITT) approach that compares subjects assigned to active treatment to subjects assigned to placebo. In an early escape design, being "rescued" is an outcome (i.e., observed after randomization) that changes the treatment actually received. For other clinical outcomes measured post-rescue, comparing the groups according to treatment assignment no longer has the clear interpretation of comparing receipt of the active treatment versus receipt of the placebo: the treatment received does not necessarily correspond to the treatment assigned. That is, the group assigned to active treatment is now composed of subjects receiving only the active treatment but also some subjects receiving both the active treatment and rescue medication. Analogously, the group assigned to placebo is now composed of some subjects receiving only placebo and some subjects receiving both placebo and rescue medication. Therefore, in an early escape design, the ITT approach is not the desired comparison between those taking the active treatment and those taking the placebo, but instead a comparison of assignment to the active treatment versus assignment to the placebo.

Early escape designs are intended to solve real ethical dilemmas (Temple, 1994; White et al., 2001). In order to address the complications that arise from early escape designs, we employ the principal stratification framework (Frangakis and Rubin, 2002), which defines latent groups based on post-treatment outcomes, here, whether

or not a subject receives rescue medication under both treatments. With principal stratification, we can define a set of subjects for which the comparison of active treatment and placebo is justified, regardless of when the outcome is measured. This approach also separates the estimand (that is, the quantity of scientific interest) from the estimation process. As we will show, an analysis that uses principal stratification differs from an analysis that only considers the subjects that are observed to be not rescued. Analyses based only on the observed outcomes can lead to estimands that do not have a causal interpretation.

In Section 1.2 we introduce data from a clinical trial of relapsing-remitting multiple sclerosis, which we use as a motivating example throughout. We introduce the principal stratification framework in Section 1.3. Comparisons to other methods are discussed in Section 1.4. In Section 1.5 we provide a formal analysis of our example via Bayesian multiple imputation. Section 1.6 concludes our discussion with extensions and relates this work to previous studies of principal stratification.

1.2 Example

Multiple sclerosis (MS) is a chronic, progressive, autoimmune disease of the central nervous system that affects a range of neurologic functions including motor, sensory, and cognitive functions. Relapsing-remitting MS is one of four subtypes and is characterized by episodes of worsening symptoms—*relapses*—followed by periods of inactivity—*remissions*. To illustrate the complications of an early escape design, we consider a placebo-controlled clinical trial of treatment for relapsing-remitting MS.

The ethical dilemma is clear: it is not in the patient's best interest to have a

chronic disease, like relapsing-remitting MS, progress while taking a placebo. Lublin and Reingold (2001), Polman et al. (2008), and Tenser (2009) discuss the ethics of using placebo-controlled or active-control trials for treatments of MS. Regardless of ethical debates, regulatory agencies (e.g., FDA) determine the requirements for clinical trials to gain approval for a new treatment. In many instances, as in the case study we present here, the guidelines expect the incorporation of a placebo arm in the trial. Early escape designs offer a compromise by allowing all patients to receive additional rescue medication if necessary.

The study was a double-blinded, multi-center, placebo-controlled randomized trial. The trial enrolled $N = 848$ subjects, with a $2:1$ allocation of active treatment to placebo. Subjects were block randomized within 99 sites and observed over two years. Several pre-treatment covariates related to disease severity were measured before treatment assignment. The covariates include the number of relapses within one year prior to treatment ($0 - 1$ vs. > 1), presence of Gadolinium (Gd) enhanced lesions detected through MRI (absent vs. present), number of T2 lesions detected through MRI (< 9 vs. ≥ 9), baseline Expanded Disability Status Scale (EDSS) score (≤ 3.5 vs. > 3.5), and age (< 40 vs. ≥ 40). Summaries of the covariates by treatment groups are provided in Table 1.1, which shows the covariates are well balanced. The primary clinical outcome of interest is the number of relapses within the first year of the study. Ideally, the treatment would reduce the the number of relapses.

Although ideally all patients would complete the study, as in most trials some patients discontinued for a variety of reasons. As such, the length of time a subject was in the study varied. For the purposes of this paper, we chose to analyze the data

at year one for several reasons. First, there was complete follow-up for all subjects up until year one. Also, estimands defined for subjects that are observed over differing lengths of time are less interpretable. Additionally, considering outcomes at one point in time provides a simpler demonstration of the principal stratification framework than considering multiple times. A more complete analysis would handle missing data due to dropouts in a more principled manner (e.g., see Barnard et al., 2003) and consider outcomes at multiple points in time.

This trial followed an early escape design that allowed subjects who did not respond to their initially assigned treatment to receive supplemental rescue medication. Subjects with some degree of progression of disability measured by EDSS were given the option to go on a pre-approved treatment in addition to the assigned study drug (either placebo or active treatment). Subjects who took the rescue medication are called "rescued". The indicator for whether or not a subject receives rescue medication within one year of first treatment is a well defined outcome.

To illustrate difficulties introduced by allowing rescue medication, consider both the primary outcome and the rescue outcome, which are summarized by treatment assignment in Table 1.1. The assignment to treatment has a significant reduction of 0.39 relapses on average. Assignment to treatment also significantly reduces the need for rescue medication by 3%. The causal effect on the receipt of rescue medication can be attributed to receipt of the active treatment versus placebo. However, because 3% of subjects assigned to active treatment and 6% of subjects assigned to placebo took additional rescue medication, the reduction of 0.39 relapses is not a comparison of receiving only active treatment versus receiving only placebo. The principal strat-

ification framework clearly defines a group of subjects for which such interpretation is correct.

Table 1.1: Averages of covariates and outcomes by treatment and placebo. All p-values come from Fisher exact tests except for relapses, which uses Welch's t test.

Covariate	**Treatment**	**Placebo**	**Difference**	**P-value**
Prior relapses	0.410	0.408	0.001	1.000
Gd lesions	0.498	0.447	0.051	0.167
T2 lesions	0.954	0.961	-0.007	0.723
EDSS	0.122	0.116	0.006	0.824
Age	0.365	0.401	-0.036	0.329
Outcome				
Rescued	0.028	0.060	-0.031	0.037
Relapses	0.245	0.637	-0.393	<0.001

1.3 Rescue medication defines principal strata

1.3.1 Potential outcomes

We work under the Rubin Causal Model (RCM; Holland, 1986), which extends Neyman's approach beyond randomized experiments and randomization-based inference (Neyman et al., 1990). For subject $i = 1, \ldots, N$, let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes of the clinical endpoint of interest under the placebo and active treatment, respectively. In early escape designs, whether or not a subject requires rescue medication is another well-defined outcome. Let $D_i(0)$ and $D_i(1)$ be the indicators of whether or not subject i received rescue medication (D for "drug") under the placebo and active treatment, respectively. In the MS trial, $Y_i(z)$ is the number of relapses within one year of starting the trial, and $D_i(z)$ is 1 if subject i receives rescue medication within one year of starting the trial and 0 otherwise, for

$z \in \{0, 1\}$. A causal effect is a comparison of the potential outcome under assignment to the active treatment versus the potential outcome under assignment to placebo for a well-defined group of subjects.

Writing the potential outcomes as functions of the indicator of assignment to treatment requires the stable unit treatment value assumption (SUTVA Rubin, 1980), which states there is no interference between subjects and no hidden versions of the treatments. In the example of the MS trial, we assume that the treatment assignment of one subject does not affect the outcome of another subject. Moreover, for each subject, there is only one version of the active treatment and the placebo.

The principal stratification framework defines "principal strata" based on post-treatment variables such as the indicator of rescue medication. The key idea is to consider the potential outcomes of the indicator of rescue medication under the two treatments jointly, but both of which are never fully observed. We encode the pair $D_i(1)$ and $D_i(0)$ more concisely as S_i:

- $S_i = \text{NN}$ if $D_i(1) = D_i(0) = 0$, subject i is *never*[1] *rescued*;
- $S_i = \text{RN}$ if $D_i(1) = 1$ and $D_i(0) = 0$, subject i is only *rescued under active treatment*;
- $S_i = \text{NR}$ if $D_i(1) = 0$ and $D_i(0) = 1$, subject i is only *rescued under placebo*;
- $S_i = \text{RR}$ if $D_i(1) = D_i(0) = 1$, subject i is *always rescued*.

We label these four principal strata the never rescued, rescued under active treatment,

[1] "Never" means regardless of treatment assignment.

rescued under placebo, and always rescued, respectively. The N notation evokes "not rescued" and R notation evokes "rescued".

With these four principal strata, the comparison of receipt of only active treatment versus receipt of only placebo is well defined solely for the NN principal stratum, which is composed of subjects who only take placebo when assigned the placebo and only take active treatment when assigned the active treatment. The average causal effect for the NN principal stratum is

$$\tau_{\mathrm{NN}} = \overline{Y}_{\mathrm{NN}}(1) - \overline{Y}_{\mathrm{NN}}(0),$$

where for $z \in \{0, 1\}$

$$\overline{Y}_{\mathrm{NN}}(z) = \frac{1}{N\pi_{\mathrm{NN}}} \sum_{i:S_i=\mathrm{NN}} Y_i(z) \quad \text{and} \quad \pi_{\mathrm{NN}} = \frac{1}{N} \sum_{i:S_i=\mathrm{NN}} 1.$$

In the example of the MS clinical trial, τ_{NN} compares the average number of relapses within one year under the active treatment (alone) and placebo (alone) for the subpopulation of subjects who would not require rescue medication under the active treatment or placebo. The never rescued principal stratum is *not* the same as the subset of units who are observed to be not rescued (Section 1.4.2 discusses this distinction in detail).

Analogous definitions of these finite population quantities apply to the other three principal strata as well. Generally, for $s \in \{\mathrm{NN}, \mathrm{RN}, \mathrm{NR}, \mathrm{RR}\}$, let τ_s be the average causal effect for principal stratum s, let $\overline{Y}_s(z)$ be the average of the potential outcomes for treatment $z \in \{0, 1\}$ in principal stratum s, and let π_s be the proportion of

subjects in principal stratum s. For subjects not in the NN stratum, the causal effect is a comparison of *assignment* to active treatment versus assignment to placebo; the interpretation is not a comparison of receiving just the active treatment versus receiving just the placebo.

The principal stratum level effects can be contrasted to the traditional finite population average causal effect, which is defined for the entire sample. The average causal effect is

$$\tau = \overline{Y}(1) - \overline{Y}(0), \tag{1.1}$$

where $\overline{Y}(z) = \frac{1}{N}\sum_{i=1}^{N} Y_i(z)$ for $z \in \{0, 1\}$.

The average causal effect will not in general reflect the principal strata level causal effects. To see why, note that the average causal effect can be rewritten as a weighted average of the principal strata level causal effects:

$$\tau = \pi_{\mathrm{NN}}\tau_{\mathrm{NN}} + \pi_{\mathrm{RN}}\tau_{\mathrm{RN}} + \pi_{\mathrm{NR}}\tau_{\mathrm{NR}} + \pi_{\mathrm{RR}}\tau_{\mathrm{RR}}.$$

In most cases, τ will differ from τ_{NN}. Moreover, neither τ nor τ_{NN} alone provide the whole picture. To support this claim, we present three hypothetical examples, which are summarized in Table 1.2. We exclude the RN stratum in these examples because we believe it is plausible to assume that the placebo cannot prevent a subject from being rescued (see Section 1.5.1 for more details). However, in complete generality the RN stratum might exist. As in the MS trial, lower values of the response are better outcomes.

Example 1.3.1 "The rescue medication dominates the active treatment". *The res-*

Table 1.2: Distinction between τ and principal strata level effects.

Example	**Principal stratum level**					**Overall**		
	s	π_s	$\overline{Y}_s(1)$	$\overline{Y}_s(0)$	τ_s	$\overline{Y}(1)$	$\overline{Y}(0)$	τ
1.3.1	RR	0.05	0.75	0.75	0.00	1.00	1.40	-0.40
	NR	0.05	1.25	0.25	1.00			
	NN	0.90	1.00	1.50	-0.50			
1.3.2	RR	0.05	2.00	4.00	-2.00	1.00	1.40	-0.40
	NR	0.05	1.50	2.50	-1.00			
	NN	0.90	0.92	1.20	-0.28			
1.3.3	RR	0.05	4.40	4.40	0.00	1.00	1.40	-0.40
	NR	0.05	3.00	3.26	-0.26			
	NN	0.90	0.70	1.13	-0.43			

cue medication is the standard of care and is actually better than the active treatment under study, which has a smaller causal effect on the outcome, but it is more expensive as well. The rescue medication under placebo helps the subjects in the NR principal stratum more than the active treatment. For subjects in the RR principal stratum, the rescue medication helps considerably under both assignments. Lastly, the active treatment has a small effect for subjects in the NN principal stratum. The average causal effects for the principal strata are $\tau_{RR} = 0$, $\tau_{NR} = 1$, *and* $\tau_{NN} = -0.5$.

Example 1.3.2 "The sicker a subject is, the more the active treatment helps". *A plausible distinction between the RR, NR, and NN principal strata is that they consist of the subjects with the worst, moderate, and best prognoses, respectively, because being rescued post treatment is an indication of disease severity. The active treatment is most effective for patients with the worst prognosis and less effective for better prognoses. Rescue medication is not helpful relative to the active treatment. The average causal effects for the principal strata are* $\tau_{RR} = -2$, $\tau_{NR} = -1$, *and* $\tau_{NN} = -0.28$.

Example 1.3.3 "A group of subjects is sick beyond help". *Again we assume that the RR, NR, and NN principal strata consist of the subjects with the worst, moderate, and best prognoses, respectively. Subjects belonging to the RR principal stratum are so sick that neither the active treatment nor the rescue medication help them. Subjects in the NR principal stratum have a slightly better prognosis, and benefit more from receiving the active treatment than the placebo and rescue medication. The treatment works best for the mildly sick subjects in the NN principal stratum. The average causal effects for the principal strata are* $\tau_{RR} = 0$, $\tau_{NR} = -0.26$, *and* $\tau_{NN} = -0.43$.

With these examples, we demonstrate that it is possible for τ to be either smaller or larger than τ_{NN}, depending on the causal effects of assignment to treatment in each principal stratum. Moreover, these examples show that considering all principal strata level effects are important for understanding the effect of treatment and its relation to receipt of additional rescue medication.

1.3.2 Observed outcomes

Of course, only one potential outcome is ever observed because each subject is only assigned to one treatment. Let Z_i be the indicator for the assignment of unit i: Z_i is 1 if subject i is assigned to the active treatment, and Z_i is 0 if subject i is assigned to placebo. The clinical endpoint that is observed is $Y_i^{\text{obs}} = Y_i(Z_i)$ and the observed indicator of receipt of rescue medication is $D_i^{\text{obs}} = D_i(Z_i)$.

In general, the indicator of treatment assignment and the observed indicator of rescue medication do not provide enough information to determine the principal stratum of each subject. To see why, consider a subject who is assigned to the active

treatment (i.e., $Z_i = 1$) and is observed to be rescued (i.e., $D_i^{\text{obs}} = 1$). We know this subject belongs either to the always rescued (RR) principal stratum or to the rescued under treatment (RN) principal stratum, but because we do not observe this subject under the placebo, we cannot determine to which of these two principal strata the subject actually belongs.

Now consider every combination of treatment assignment and observed indicator of receipt of rescue medication. We denote

$$\mathcal{O}(z, d) = \{i : Z_i = z \text{ and } D_i^{\text{obs}} = d\}$$

for $z \in \{0, 1\}$ and $d \in \{0, 1\}$. The $\mathcal{O}$ notation evokes "observed groups". The possible membership to certain principal strata for each of the four combinations of treatment assignment and observed indicator of rescue medication are detailed in Table 1.3. Although the principal stratum of any one individual is not identified in general, the principal stratification framework brings clarity to the idea that the observed indicator of rescue medication is not enough information to characterize whether or not a subject takes only the active treatment and only the placebo. Ignoring or misusing the observed indicator of rescue medication leads to misleading conclusions about the effect of taking the active treatment compared to taking the placebo.

Table 1.3: Observed and latent groups.

Assignment	**Rescued**	$\mathcal{O}(Z_i, D_i^{\text{obs}})$	S_i
Active treatment	Yes	$\mathcal{O}(1, 1)$	RR or RN
	No	$\mathcal{O}(1, 0)$	NR or NN
Placebo	Yes	$\mathcal{O}(0, 1)$	RR or NR
	No	$\mathcal{O}(0, 0)$	NN or RN

1.4 Other methods of analysis

We review some methods for analyzing data from clinical trials with early escape designs. An important characteristic of the principal stratification approach is that it starts by defining the estimand, that is, the quantity of scientific interest. Separating the estimand from estimation keeps the goal of the study clear. We assert that this separation is one of the main distinctions between the principal stratification framework and other methods of analysis that focus on estimators.

1.4.1 Intention-to-treat analysis

As discussed, ITT analysis compares all subjects assigned to the active treatment to all subjects assigned to the placebo regardless of the treatments actually received. The difference of the observed means in the treatment and control groups is an unbiased ITT estimator of the average causal effect, τ as defined in (1.1). In clinical trials, randomization validates the ITT analysis, but ITT answers questions about causal effects of *assignment* to active treatment versus assignment to placebo. However, in an early escape design, ITT does not address the effects of *receipt* of only active treatment versus only placebo.

1.4.2 Discarding rescued subjects

An analysis that *discards rescued subjects* removes all subjects who were rescued. One such estimator is the difference in means of subjects assigned to the active treat-

ment versus placebo for subjects observed to be not rescued. That is,

$$\hat{\phi}_{\text{dr}} = \frac{\sum_{i \in \mathcal{O}(1,0)} Y_i^{\text{obs}}}{\sum_{i \in \mathcal{O}(1,0)} 1} - \frac{\sum_{i \in \mathcal{O}(0,0)} Y_i^{\text{obs}}}{\sum_{i \in \mathcal{O}(0,0)} 1}, \tag{1.2}$$

is an estimator that discards rescued subjects, but it is not clear what $\hat{\phi}_{\text{dr}}$ estimates. The "dr" subscript evokes "discard rescued".

The problem with this approach is that it only discards subjects that were rescued under the observed treatment assignment. Consequently, the two groups being compared are not a common set of subjects (see Table 1.3). The subjects who are assigned to the active treatment and are observed to be not rescued belong to either the rescued under placebo (NR) or the never rescued (NN) principal strata. The subjects who are assigned to the placebo and are observed to be not rescued belong to either the rescued under treatment (RN) or the never rescued (NN) principal strata. An analysis that discards rescued subjects does not estimate a causal effect because it is a comparison of potential outcomes between two dissimilar groups.

To see this issue analytically, we see that $\hat{\phi}_{\text{dr}}$ estimates ϕ_{dr} where

$$\begin{aligned}
\phi_{\text{dr}} =& \frac{\pi_{\text{NN}}\overline{Y}_{\text{NN}}(1) + \pi_{\text{NR}}\overline{Y}_{\text{NR}}(1)}{\pi_{\text{NN}} + \pi_{\text{NR}}} - \frac{\pi_{\text{NN}}\overline{Y}_{\text{NN}}(0) + \pi_{\text{RN}}\overline{Y}_{\text{RN}}(0)}{\pi_{\text{NN}} + \pi_{\text{RN}}} \\
=& \tau_{\text{NN}} + \frac{\pi_{\text{NR}}}{\pi_{\text{NN}} + \pi_{\text{NR}}}\big(\overline{Y}_{\text{NR}}(1) - \overline{Y}_{\text{NN}}(1)\big) + \frac{\pi_{\text{RN}}}{\pi_{\text{NN}} + \pi_{\text{RN}}}\big(\overline{Y}_{\text{NN}}(0) - \overline{Y}_{\text{RN}}(0)\big).
\end{aligned}$$

By "estimates", we mean $\hat{\phi}_{\text{dr}}$ is a biased estimate of ϕ_{dr} under repeated randomization of assignment to treatment with a bias that goes to zero as N increases (details are in Appendix A.1). Clearly, this estimand is not in general the same as τ_{NN}.

One case in which ϕ_{dr} equals τ_{NN} is when $\pi_{\text{NR}} = \pi_{\text{RN}} = 0$. This case corresponds to the testable assumption that the active treatment has no effect on the receipt of rescue medication, or that $D_i(0) = D_i(1)$ for all $i = 1, \ldots, N$. In our case study, the small p-value from the "Rescued" row of Table 1.1 suggests that this assumption does not hold.

1.4.3 Rescue as a "bad" outcome

White et al. (2001) remark that being rescued can be perceived as a bad prognostic outcome. Indeed, in the MS clinical trial, the decision to give subjects rescue medication was based on progression of the disease. In such a case, White et al. (2001) suggest imputing a "bad underlying outcome to rescued patients and use rank-based methods of analysis". The goal of such an approach is to estimate the effect of the active treatment compared to the placebo for the entire sample under a hypothetical alternative experiment that had no early escape protocol (i.e., the *one versus none* comparison from White et al., 2003).

This description would benefit from expanding the notation of the potential outcomes to allow the outcomes to be a function of receipt of rescue medication. We expand the notation for just this section. Let $Y_i(z, d)$ be the potential outcome where $z \in \{0, 1\}$ indicates which treatment a subject is assigned to, and $d \in \{0, 1\}$ indicates whether or not a subject receives rescue medication. The one versus none comparison is an estimand that compares $Y_i(1, 0)$ to $Y_i(0, 0)$ for all subjects.

Expanding the potential outcomes changes the problem from defining different underlying principal strata to considering the causal effect of receipt of rescue med-

ication. A crucial component of the RCM is the assignment mechanism, which is unknown for the receipt of rescue medication because it is based on physicians' subjective opinions of the putative (or possibly observed) outcome of interest. An unknown assignment mechanism means that rescue medication is not under the control of the experimenter and puts this causal inference question in the domain of observational studies.

White et al. (2001) attempt to obviate this problem by imputing a "bad" potential outcome for rescued subjects and using a rank-based analysis. Using rank-based methods implicitly changes the estimand of interest. For example, after imputing the bad outcomes, if the median of the subjects assigned to active treatment is compared to the median of the subjects assigned to placebo, then the estimand is the difference in medians of the potential outcomes under active treatment and placebo with no rescue medication; that is, we estimate

$$\text{median}\{Y_i(1,0) : i = 1, \ldots, N\} - \text{median}\{Y_i(0,0) : i = 1, \ldots, N\}.$$

The difference in medians is a reasonable estimand, but such an estimand should be scientifically motivated (e.g., when the outcome distribution is believed to be skewed). The early escape design should not dictate the choice of estimand.

1.5 Parametric estimation for MS clinical trial

We conduct a Bayesian parametric analysis of the MS clinical trial described in Section 1.2. The goal is to estimate τ_{NN}, the average causal effect for the never

rescued stratum. We also estimate the average causal effects in the other principal strata. Additional estimands include the proportion of subjects within each stratum, the average potential outcomes by treatment, and the overall average causal effect.

We view the unobserved principal strata membership and the missing potential outcome for the clinical endpoint for each subject as missing data. Our estimation strategy involves multiply imputing all the missing data using a Bayesian model (Rubin, 2009). First we impose structural and stochastic assumptions on the principal strata and the potential outcomes models, respectively. Then we provide explicit parametric models for the complete data, which include the indicator of treatment assignment, principal strata membership, and both potential outcomes for the clinical endpoint. The latent strata and missing potential outcomes are then imputed from their posterior predictive distribution given the observed data. Our estimands are simple functions of the completed data, thus every imputation provides a posterior draw of the estimands.

1.5.1 Assumptions

Monotonicity

We assume monotonicity on the principal strata. In terms of potential outcomes notation, this assumption implies $D_i(1) \leq D_i(0)$ for all subjects i. Intuitively, the interpretation is that no subject would require rescue medication under the active treatment but not under the placebo. Monotonicity *a priori* precludes the existence of the RN principal stratum. This assumption is credible for receipt of rescue medication in a placebo-controlled clinical trial because the decision to administer rescue

medication is based on how well the subject responds to the assigned treatment, and it is unlikely that the active treatment is worse than the placebo. For our example, if a subject requires rescue medication under the treatment that is suspected to slow down MS progression, then it is plausible that this subject would also require rescue medication under the placebo, which is known to have no biological effect. Similarly, if a subject does not require rescue medication under the inactive placebo, then this subject would not require rescue medication under the active treatment, which is not believed to have negative effects on MS progression. In other trials, this assumption may not be as plausible.

Stochastic dominance

We assume a stochastic ordering on the potential outcomes of the clinical endpoint between certain principal strata. Namely, the assumptions follow from the argument that the RR, NR, and NN principal strata correspond to the worst, moderate, and best prognostic groups. Thus, the NN principal stratum should have better outcomes than the NR principal stratum under the active treatment because subjects in both strata receive only the active treatment when assigned the active treatment. We encode this assumption through stochastic dominance of the potential outcome under active treatment in the NR principal stratum over the NN principal stratum (recalling that a lower number of relapses is better). Similarly, the NR principal stratum should have better outcomes than the RR principal stratum under the placebo because subjects in both strata receive placebo and rescue medication when assigned the placebo. We encode this assumption through stochastic dominance of the potential outcome under

the placebo in the RR principal stratum over the NR principal stratum.

Stochastic dominance is an assumption on the distribution of potential outcomes. More explicitly, our first stochastic dominance assumption implies that for subject i in the NN principal stratum and subject i' in the NR principal stratum,

$$\Pr(Y_{i'}(1) \leq y \mid S_{i'} = \text{NR}, \boldsymbol{\theta}) \leq \Pr(Y_i(1) \leq y \mid S_i = \text{NN}, \boldsymbol{\theta})$$

for all possible y and $\boldsymbol{\theta}$ is a vector parameter governing the distributions. A similar definition applies to our second stochastic dominance assumption. Later, we make both of these assumptions explicit with respect to a parametric model and conditional on covariates.

1.5.2 Complete data model

We start with a model for the complete data given pre-treatment covariates and parameters. For subject i, the complete data are composed of the indicator of treatment assignment, Z_i; the potential outcomes for the primary endpoint under assignment to placebo and active treatment, $Y_i(0)$ and $Y_i(1)$ respectively; and the principal strata membership S_i (or equivalently, both potential outcomes of the indicator of rescue under assignment to placebo and active treatment, $D_i(0)$ and $D_i(1)$ respectively). We denote the indicator for site (or hospital) for each unit as the 99-dimensional column vector $\boldsymbol{X}_i^{\text{h}}$, where "h" evokes "hospital". We denote the five pre-treatment covariates introduced in Section 1.2 with the five-dimensional column vector $\boldsymbol{X}_i$, and the vector of parameters as $\boldsymbol{\theta}$. The five binary covariates listed in Table 1.1 are believed to be predictive of the outcome of number of relapses and are coded as -1 and

1. The complete data likelihood factorizes as follows:

$$\begin{aligned} p(Z_i, Y_i(0), Y_i(1), S_i \mid \boldsymbol{X}_i^{\mathrm{h}}, \boldsymbol{X}_i, \boldsymbol{\theta}) &= p(Z_i \mid Y_i(0), Y_i(1), S_i, \boldsymbol{X}_i^{\mathrm{h}}, \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{Z}}) \\ &\times p(Y_i(0), Y_i(1) \mid S_i, \boldsymbol{X}_i^{\mathrm{h}}, \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{Y}}) \\ &\times p(S_i \mid \boldsymbol{X}_i^{\mathrm{h}}, \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{S}}), \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathrm{Z}}, \boldsymbol{\theta}^{\mathrm{S}}, \boldsymbol{\theta}^{\mathrm{Y}})$. We consider each term individually.

Treatment assignment model

As discussed in Section 1.2, the trial was randomized within site (or hospital). Thus, given the site, the treatment assignment indicator is independent of everything else: assignment to treatment is unconfounded. That is,

$$p(Z_i \mid Y_i(0), Y_i(1), S_i, \boldsymbol{X}_i^{\mathrm{h}}, \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{Z}}) = p(Z_i \mid \boldsymbol{X}_i^{\mathrm{h}}).$$

Treatment assignment does not depend on any unknown parameters.

Principal strata model

With the monotonicity assumption, subjects belong to one of three principal strata. We model the probability of membership to the NN principal stratum (i.e., the healthiest group) versus the NR and RR principal strata (i.e., the unhealthier groups) with a logistic regression:

$$\pi_{\mathrm{NN},i} = \Pr(S_i = \mathrm{NN} \mid \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{S}}) = \mathrm{expit}(\beta_{\mathrm{NN}} + \boldsymbol{X}_i^{\top} \boldsymbol{\beta}_{\mathrm{NN,X}}).$$

where $\text{expit}(l) = 1/(1 + e^{-l})$. Given that a subject does not belong to the NN principal stratum, the conditional probability of belonging to the NR principal stratum (i.e., moderately healthy group) versus the RR principal strata (i.e., the unhealthiest group) is also modeled with a logistic regression:

$$\pi_{\text{RN}|\neg\text{NN},i} = \Pr(S_i = \text{NR} \mid S_i \in \{\text{NR or RR}\}, \boldsymbol{X}_i, \boldsymbol{\theta}^{\text{S}}) = \text{expit}(\beta_{\text{NR}} + \boldsymbol{X}_i^{\top}\boldsymbol{\beta}_{\text{NR,X}})$$

The conditional nature of β_{NR} and $\boldsymbol{\beta}_{\text{NR,X}}$ is suppressed for notational convenience. The unconditional probability of belonging the NR principal strata is

$$\pi_{\text{NR},i} = \Pr(S_i = \text{NR} \mid \boldsymbol{X}_i, \boldsymbol{\theta}^{\text{S}}) = \pi_{\text{RN}|\neg\text{NN},i} \cdot (1 - \pi_{\text{NN},i})$$

We have $\boldsymbol{\theta}^{\text{S}} = (\beta_{\text{NN}}, \beta_{\text{NR}}, \boldsymbol{\beta}_{\text{NN,X}}, \boldsymbol{\beta}_{\text{NR,X}})$.

We do not include information on site in this model and therefore suppress $\boldsymbol{X}_i^{\text{h}}$ from the notation. The implicit assumption is that the covariates, $\boldsymbol{X}_i$, provide enough information to ignore possible differences across the 99 sites. A more complete analysis would incorporate possible differences in sites with a hierarchical model (Gelman et al., 2003, Chapter 5).

Primary endpoint model

We model the number of relapses with a Poisson distribution, which is commonly employed in MS research. This model is conditional on both covariates and principal strata membership. As in the principal strata model, we suppress $\boldsymbol{X}_i^{\text{h}}$ from the notation. For every treatment assignment and principal stratum combination, we

marginally model the potential outcome using a Poisson regression:

$$Y_i(z) \mid S_i = s, \boldsymbol{X}_i, \boldsymbol{\theta}^{\mathrm{Y}} \sim \mathrm{Poi}(\lambda_{z,s,i}),$$

where

$$\log \lambda_{z,s,i} = \gamma_{z,s} + \boldsymbol{X}_i^{\top} \boldsymbol{\gamma}_{z,s,\mathrm{X}}$$

and $z \in \{0, 1\}$ and $s \in \{\mathrm{RR}, \mathrm{NR}, \mathrm{NN}\}$. We simplify this general model to have constant coefficients for the covariates, so $\boldsymbol{\gamma}_{z,s,\mathrm{X}} = \boldsymbol{\gamma}_{\mathrm{X}}$ for all assigned treatment and principal stratum combinations. We have $\boldsymbol{\theta}^{\mathrm{Y}} = (\gamma_{0,\mathrm{RR}}, \gamma_{0,\mathrm{NR}}, \gamma_{0,\mathrm{NN}}, \gamma_{1,\mathrm{RR}}, \gamma_{1,\mathrm{NR}}, \gamma_{1,\mathrm{NN}}, \boldsymbol{\gamma}_{\mathrm{X}})$. We assume conditional independence between $Y_i(0)$ and $Y_i(1)$.

The stochastic dominance assumption manifests in terms of inequalities on the parameters of the outcome model. First, stochastic dominance of the NR principal stratum over the NN principal stratum for the potential outcome under active treatment implies $\gamma_{1,\mathrm{NR}} \geq \gamma_{1,\mathrm{NN}}$. Second, stochastic dominance of the RR principal stratum over the NR principal stratum for the potential outcome under placebo implies $\gamma_{0,\mathrm{RR}} \geq \gamma_{0,\mathrm{NR}}$.

1.5.3 Imputation of missing data

Observed data likelihood

Both potential outcomes are never jointly observed, so we work from the observed data likelihood, which integrates over the missing data. The observed likelihood is decomposed into four combinations of treatment assignment and observed rescue status. Under monotonicity, subjects that are rescued under the active treatment

belong to the always rescued principal stratum because they would have also been rescued under the placebo. Likewise, subjects that are not rescued under placebo belong to the never rescued principal stratum because they would not require rescue medication under the active treatment. In contrast, the principal stratum of subjects rescued under placebo cannot be uniquely identified from the observed data because they could have been either rescued or not rescued under the active treatment. These subjects contribute a Poisson finite mixture probability with two components to the likelihood. A similar argument applies to subjects that are not rescued under the active treatment because they could have been either rescued or not rescued under the placebo. These subjects also contribute a Poisson finite mixture probability with two components to the likelihood. The observed data likelihood is

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{X}) \quad \propto \quad & \prod_{i \in \mathcal{O}(1,1)} \pi_{\text{RR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{1,\text{RR},i}) \\
\times \quad & \prod_{i \in \mathcal{O}(1,0)} \pi_{\text{NN},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{1,\text{NN},i}) + \pi_{\text{NR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{1,\text{NR},i}) \\
\times \quad & \prod_{i \in \mathcal{O}(0,1)} \pi_{\text{NR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{NR},i}) + \pi_{\text{RR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{RR},i}) \\
\times \quad & \prod_{i \in \mathcal{O}(0,0)} \pi_{\text{NN},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{NN},i}) \\
\times \quad & \mathbb{1}\{\gamma_{1,\text{NR}} \geq \gamma_{1,\text{NN}}\} \mathbb{1}\{\gamma_{0,\text{RR}} \geq \gamma_{0,\text{NR}}\},
\end{aligned}
$$

where $Poi(y; \lambda)$ is the probability mass function of the Poisson distribution with mean $\lambda > 0$ evaluated at $y = 0, 1, \ldots$; $\mathbf{Z}$, $\mathbf{Y}^{\text{obs}}$, and $\mathbf{D}^{\text{obs}}$ are vectors composed of the values Z_i, Y_i^{obs}, and D_i^{obs} for $i = 1, \ldots, N$ respectively; and $\mathbf{X}$ is an $N \times 5$ matrix with rows $\boldsymbol{X}_i^\top$ for $i = 1, \ldots, N$. The last line is a product of indicator functions that enforces the stochastic dominance assumption.

Prior distributions

We use informative proper prior distributions for the parameters in order to ensure a proper posterior distribution. Our choice of hyperparameters induces a prior distribution on the finite population parameters, which are summarized in Table 1.4. All prior distributions are relatively diffuse and cover a plausible range of values. All causal effects have medians near or at zero.

We assume normal prior distributions for all parameters. The prior distributions for $\boldsymbol{\beta}_{\mathrm{RR,X}}$ and $\boldsymbol{\beta}_{\mathrm{NR,X}}$ are independent normal distributions with mean 0 and variance 1 for each entry. The prior distributions for β_{RR} and β_{NR} are independent normal distributions with mean -2.89 and variance 1. The prior distributions for $\boldsymbol{\gamma}_{\mathrm{X}}$ are independent normal distributions with mean 0 and variance 1 for each entry.

The prior distributions on the remaining parameters, $\gamma_{z,s}$, were chosen to ensure three characteristics: (1) we target a median of $\overline{Y}_{\mathrm{NN}}(0)$ of approximately 0.5 based on the results in the placebo arms of similar clinical trials (Gold et al., 2013; Polman et al., 2006; Gold et al., 2012; Fox et al., 2012); (2) we require our two stochastic dominance assumptions, that is $\gamma_{1,\mathrm{NR}} \geq \gamma_{1,\mathrm{NN}}$ and $\gamma_{0,\mathrm{RR}} \geq \gamma_{0,\mathrm{NR}}$, and (3) we target an "approximately null" distribution of causal effects by setting $\mathrm{E}(\gamma_{0,\mathrm{RR}}) = \mathrm{E}(\gamma_{1,\mathrm{RR}})$, $\mathrm{E}(\gamma_{0,\mathrm{NR}}) = \mathrm{E}(\gamma_{1,\mathrm{NR}})$, and $\mathrm{E}(\gamma_{0,\mathrm{NN}}) = \mathrm{E}(\gamma_{1,\mathrm{NN}})$. Specifically, $\gamma_{1,\mathrm{RR}}$ is distributed as a normal distribution with mean 0.16 and variance 1; $\gamma_{1,\mathrm{NR}}$ and $\gamma_{1,\mathrm{NN}}$ are distributed as the maximum and minimum respectively of two independent normal random variables with mean -1.53 and variance 1; $\gamma_{0,\mathrm{NN}}$ is distributed as a normal distribution with mean -2.09 and variance 1; and $\gamma_{0,\mathrm{RR}}$ and $\gamma_{0,\mathrm{NR}}$ are distributed as the maximum and minimum respectively of two independent normal random variables with mean -0.4

and variance 1. Characteristic (3) is guaranteed by the moments of the maximum of two normal random variables (Clark, 1961), (2) is guaranteed by the imposed orderings, and (1) is confirmed in Table 1.4.

Table 1.4: Summaries of the prior distributions of several estimands. Outcome values are right skewed, so the median and quantiles provide better summaries of centrality and spread than the mean and standard deviation, respectively.

Estimand	**Mean**	**Std. Dev.**	**Median**	**2.5 Percentile**	**97.5 Percentile**
π_{RR}	0.116	0.11	0.078	0.006	0.421
π_{NR}	0.116	0.11	0.079	0.006	0.423
π_{NN}	0.768	0.17	0.812	0.342	0.980
$\overline{Y}_{\text{RR}}(1)$	24.453	192.02	3.601	0.092	155.230
$\overline{Y}_{\text{NR}}(1)$	7.595	59.58	1.143	0.014	40.106
$\overline{Y}_{\text{NN}}(1)$	2.105	8.87	0.474	0.031	13.273
$\overline{Y}(1)$	5.551	50.076	1.09	0.086	32.014
$\overline{Y}_{\text{RR}}(0)$	25.093	262.34	3.538	0.125	143.310
$\overline{Y}_{\text{NR}}(0)$	7.304	65.82	1.144	0.016	43.568
$\overline{Y}_{\text{NN}}(0)$	2.687	18.95	0.469	0.025	15.792
$\overline{Y}(0)$	5.766	48.45	1.091	0.081	33.023
τ_{RR}	-0.640	202.68	0.000	-71.761	83.521
τ_{NR}	0.291	44.39	0.000	-20.597	20.305
τ_{NN}	-0.582	16.23	0.007	-8.816	6.410
τ	-0.215	41.06	0.001	-13.527	13.433

Posterior sampling

Posterior samples of the parameters were drawn using STAN software (Stan Development Team, 2013). STAN implements the No-U-Turn sampler (Hoffman and Gelman, 2011), a variant of Hamiltonian Monte Carlo (Duane et al., 1987). Ten chains of 20,000 iterations were run simultaneously in order to assess convergence. All potential scale reduction factors were below 1.001 for all parameters (Gelman and Rubin, 1992). No stochastic dominance assumptions were enforced while sampling the posterior distribution. A sample of parameters following stochastic dominance

was obtained through rejection sampling. That is, we only use the posterior samples for which both stochastic dominance assumptions are true. Software was verified using a simulation based procedure (results in Appendix A.3, see Cook et al., 2006, for details on the procedure). Model fit was assessed using posterior predictive checks (results in Appendix A.4, see Rubin et al., 1984, for details on the procedure).

Imputing missing potential outcomes

The observed data include the observed indicator of receipt of rescue medication and the observed clinical outcome, which constitute half of the potential outcomes. We multiply impute the missing potential outcomes given the observed data and a posterior sample of the parameters (Rubin, 2009). First, we impute principal stratum membership for the subjects who could belong to two possible principal strata. Next, given the imputed principal stratum for each subject (and the posterior sample of the parameters), we can impute the missing potential outcome of number of relapses within one year under the alternative treatment. The details for the imputation are provided in Appendix A.2. Once all missing data are imputed, we have a posterior sample of the complete data, from which calculating all estimands of interest is trivial. Each fully imputed dataset—or completed dataset—leads to one posterior draw of the estimands given the observed data.

There is a non-zero probability of imputing no subjects that belong to the NR principal stratum. We assume *a priori* that this principal stratum exists because the treatment appears to have an effect on receipt of rescue medication (see Table 1.1). Based on this assumption, we exclude the posterior samples of the completed data

that had no subjects in the NR principal stratum (0.4% of the posterior sample).

Figure 1.1 shows the posterior distribution of the three principal strata level effects; the posterior distribution is superimposed over the prior distribution for easy comparison. We see that the posterior distribution of the average causal effect for the never rescued principal stratum has a smaller variance relative to its prior distribution. The same is true for the always rescued principal stratum. The posterior distribution of the average causal effect for the NR principal stratum is less concentrated than the other two principal strata. Even with a parametric model and stochastic dominance assumptions, estimating the effects in this group is difficult. The general reason is that no subjects are ever completely identified to belong to this stratum, whereas with the monotonicity assumption, subjects in observed groups $\mathcal{O}(1,1)$ and $\mathcal{O}(0,0)$ are identified to be in the always rescued and never rescued principal strata respectively. Moreover, for this particular problem, the rescued under placebo principal stratum is composed of approximately 3% of the sample (or about 25 subjects), which is a small number of observations for estimating the log means of the potential outcomes (i.e., $\lambda_{0,\text{NR}}$ and $\lambda_{1,\text{NR}}$) within two different two component mixture models.

1.5.4 Results

Summaries of the posterior distributions of all estimands of interest are provided in Table 1.5. Figures 1.1 and 1.2 show the histograms of the posterior (and prior) distributions of the average causal effects within strata and overall, respectively.

We estimate that 3.2% and 2.4% of the sample belong to the always rescued and rescued under placebo principal strata respectively. The remaining 94.4% belong to

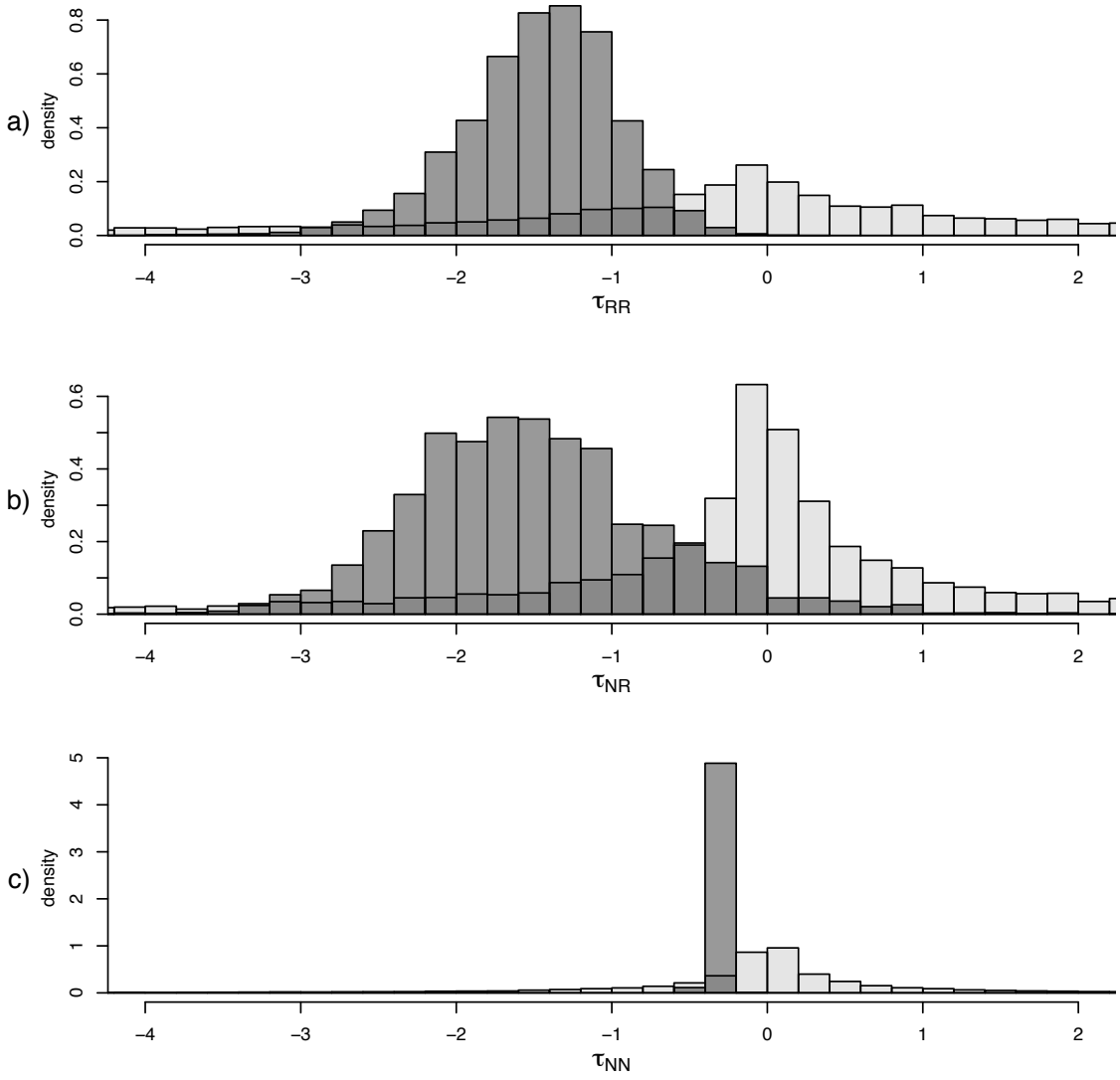

Figure 1.1: Posterior (and prior) distributions of the average causal effects within the a) always rescued, b) rescued under placebo, and c) never rescued principal strata. Posterior distributions are plotted in dark grey and prior distributions are plotted in light grey.

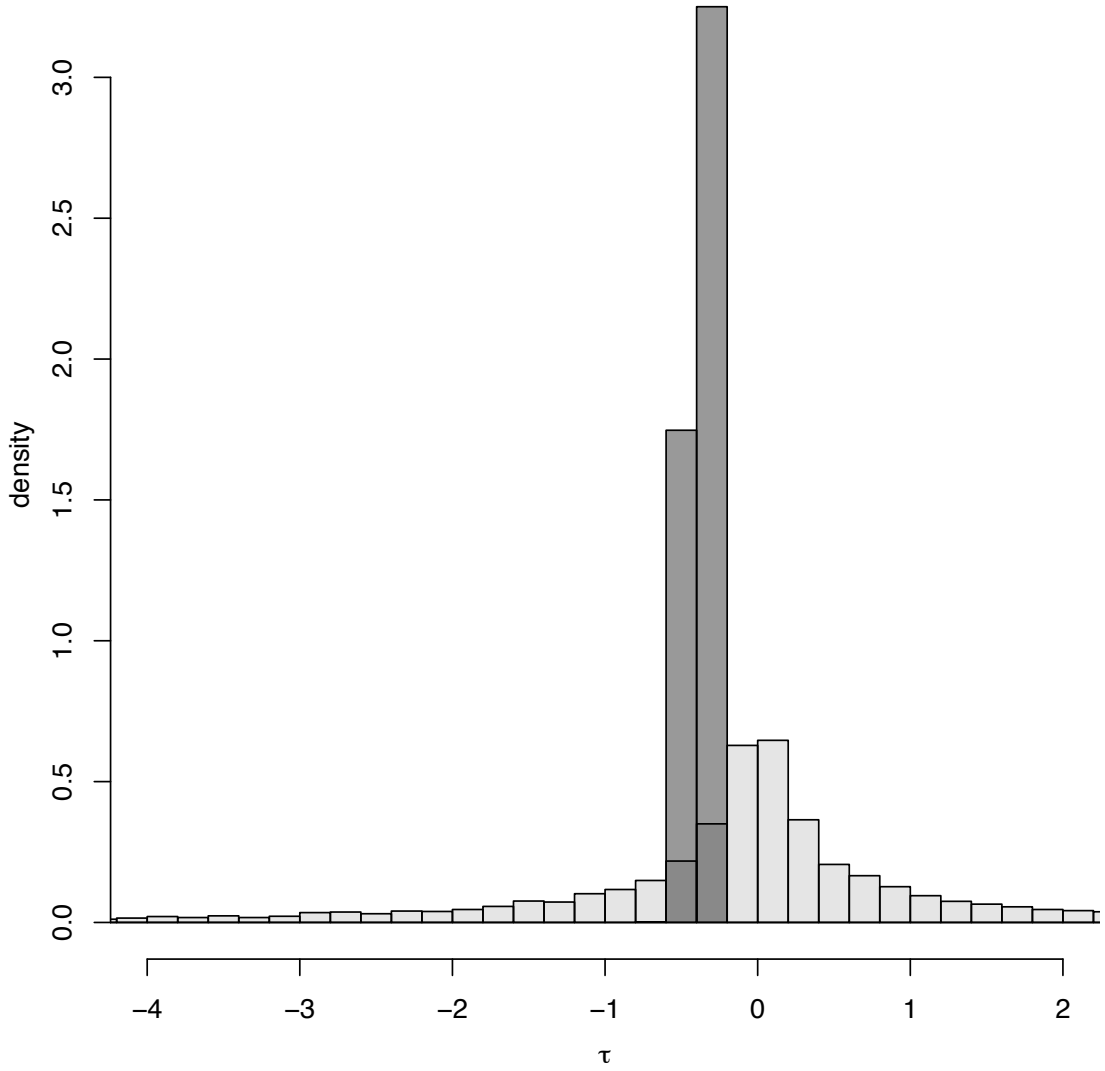

Figure 1.2: Posterior (and prior) distribution(s) of the overall average causal effect. Posterior distribution plotted in dark grey and prior distribution plotted in light grey.

Table 1.5: Summaries of the posterior distributions of several estimands.

Estimand	**Mean**	**Std. Dev.**	**Median**	**2.5$^{\text{th}}$ Percentile**	**97.5$^{\text{th}}$**
π_{RR}	0.032	0.004	0.032	0.024	0.039
π_{NR}	0.024	0.013	0.022	0.004	0.053
π_{NN}	0.944	0.010	0.946	0.921	0.959
$\overline{Y}_{\text{RR}}(1)$	1.408	0.188	1.393	1.077	1.812
$\overline{Y}_{\text{NR}}(1)$	0.541	0.362	0.469	0.000	1.400
$\overline{Y}_{\text{NN}}(1)$	0.199	0.014	0.200	0.171	0.225
$\overline{Y}(1)$	0.246	0.013	0.245	0.222	0.274
$\overline{Y}_{\text{RR}}(0)$	2.864	0.459	2.818	2.120	3.900
$\overline{Y}_{\text{NR}}(0)$	2.021	0.707	2.065	0.444	3.308
$\overline{Y}_{\text{NN}}(0)$	0.515	0.037	0.514	0.446	0.591
$\overline{Y}(0)$	0.630	0.045	0.626	0.548	0.724
τ_{RR}	-1.456	0.504	-1.417	-2.560	-0.586
τ_{NR}	-1.480	0.805	-1.550	-2.889	0.300
τ_{NN}	-0.316	0.039	-0.315	-0.397	-0.242
τ	-0.384	0.049	-0.382	-0.485	-0.295

the never rescued principal stratum. The overall causal average effect of assignment to treatment, τ, is a reduction of 0.38 relapses on average.

Within the never rescued principal stratum, our inference shows that receipt of only the active treatment reduces the number of relapses in one year compared to receipt of only the placebo. The estimated effect is an average reduction of 0.32 relapses. Additionally, our analysis shows that assignment to treatment is still effective, and in fact more effective, for subjects in the always rescued and rescued under placebo principal strata; the estimated average causal effects for these strata are a reduction of 1.46 and 1.48 relapses, respectively. These estimates suggest an explanation similar to "the sicker a subject is, the more the active treatment helps" example because the average effects for the never rescued and rescued under placebo principal strata are larger in magnitude than the average effect for never rescued principal stratum.

Estimation of the principal strata level effects shows that receipt of treatment

is indeed more effective than receipt of placebo within the never rescued group of patients. Our analysis additionally characterizes the causal effects in the other two principal strata. Although it is hard to disentangle the possible effects of the rescue medication, the assignment to treatment still benefits subjects requiring some rescue medication.

1.6 Conclusion

We described the problem of administering rescue medication after treatment has been randomized in a clinical trial. The principal stratification framework clarifies the difficulties of these early escape designs as well as defines a group for which the desired comparison of receipt of only active treatment and receipt of only placebo is appropriate.

There are several extensions of this work that might occur in practice. For instance, we have ignored the temporal aspect of when rescue medications are actually given to the patient. We ignore this issue in our example for two reasons: first, the binary indicator of receipt of rescue medication is easier to understand, and second, the additional complication of dealing with continuous principal strata is probably not required for such a small number of rescued patients. If an early escape design took place over a long period of time with a larger proportion of observed patients requiring rescue medication (e.g., see the example in White et al., 2001), then considering the principal strata of time of rescue under active treatment and control would be an important extension. Such extensions could follow similar work on partial compliance (e.g., Jin and Rubin, 2008).

Another application could be to active-control trials, or non-inferiority trials. Although the ethical issues are of less concern, early escape designs are still appropriate because patients that do not respond to the assigned treatment may benefit from additional rescue medication. An additional complication might be the violation of some of our assumptions. For example, the monotonicity assumption seems reasonable in a placebo-controlled trial because the placebo is inactive, but with an active-control trial, we can envision that some subjects might benefit more under the control treatment than the experimental treatment and vice versa.

We detailed a parametric Bayesian inference with strong, plausible assumptions to estimate the different principal strata level effects. Other estimation methods have been proposed, including large sample bounds (Balke and Pearl, 1997; Zhang and Rubin, 2003), sensitivity analysis (Gilbert et al., 2003; Hudgens and Halloran, 2006), and randomization-based inference (Rosenbaum, 1996; Nolen and Hudgens, 2011). Future work might apply these approaches to estimation of causal effects in early escape designs.

Principal stratification is a general framework that describes latent groups based on post-treatment variables under both treatments. Causal estimands are well defined for these principal strata. In contrast, methods that rely only on the observed outcomes—like analyses that discard rescued subjects—fail to consider what would have happened under the alternative treatment. Similar observations apply to "per-protocol" and "as-treated" analyses of randomized experiments with imperfect compliance (Angrist et al., 1996).

The principal stratification framework also addresses the issue of non-compliance

in another ethically motivated design, the encouragement design (Hirano et al., 2000), in which subjects are randomly assigned to either be encouraged to take a treatment (e.g., a mailing that reminds subjects to get an influenza vaccine) or not encouraged (e.g., no mailing). The treatment is not withheld from individuals on ethical grounds. Perhaps knowledge of how to handle conceptually post-treatment variables can motivate designs that address other ethical concerns.

We provided an example of estimation of principal strata level effects for subjects that receive additional rescue medication in early escape designs. Previously, early escape designs were thought to be only useful for determining effects of treatment versus placebo up until the first receipt of rescue medication (Temple, 1994; FDA, 2001). However, we showed that we can estimate the causal effect of receipt of active treatment compared to receipt of placebo for the never rescued principal stratum. Moreover, within the other principal strata, the effects of assignment to treatment are important for understanding the effect of the treatment even when rescue medication is required. We hope that our example and the general framework provide evidence for the further use of early escape designs that alleviate the ethical dilemma of treating patients with an inactive placebo.

Chapter 2

Randomization-based intervals for binary outcomes

2.1 Introduction

Fisher (1935, page 17) called randomization in experiments the "physical basis of the validity" of the randomization test that assumes the sharp null hypothesis of absolutely no treatment effect. Fisher's randomization test is appealing because it provides simple, non-parametric inference for assessing causal effects (Pitman, 1937a,b, 1938). Rejecting the null hypothesis leads to the "dull" conclusion that there is at least one unit for which the treatment does have an effect. Although inverting sets of hypothesis tests of constant additive effects provides interval estimates of the average causal effect, with binary outcomes, non-zero additive effects either are -1 or 1 and are generally contradicted by observed data. We formally treat the problem of testing a null hypotheses for the average causal effect in the case of binary outcomes using the

randomization employed in the experiment as our primary inferential tool. Inverting such tests provide interval estimates of the average causal effect.

For binary outcomes, a hypothesis with respect to the average causal effect can be defined, but necessarily requires the introduction of nuisance unknowns that complicate the testing procedure. For example, if the null hypothesis asserts that the treatment protects against disease for exactly one unit, we cannot know if that one unit was observed to be a non-diseased person assigned to treatment or a diseased person assigned to control, and thus we are unable to fill-in the missing potential outcomes to continue with a randomization test. We propose the interpretation of Fisher's randomization test as a special case of a posterior predictive check as way to test null hypotheses of average causal effects (Rubin et al., 1984).

In Section 2.2, we describe the situation with a completely randomized experiment. After some background and review, the problem of nuisance unknowns is demonstrated with a simple heuristic and then generalized. In section 2.3 we suggest a solution using posterior predictive checks and discuss other approaches. Section 2.4 provides toy examples including an application to a small dataset and evaluation of repeated sampling frequentist properties. In Section 2.5 we discuss extensions of our procedure as well as limitations.

2.2 Analysis of a completely randomized experiment

2.2.1 Background

The Rubin Causal Model (Rubin, 1978a; Holland, 1986) forms the basis of our entire discussion. Consider comparing a binary outcome under an active treatment and control treatment for N units. For concreteness, consider outcome 1 as "diseased" and 0 as "non-diseased". We now denote the potential outcomes for unit i under control and active treatments as $Y_i(0)$ and $Y_i(1)$, respectively. Let $\mathbf{Y}(w)^\top = (Y_1(w), ..., Y_N(w))^\top$ for $w = 0, 1$; $\mathbf{Y} = [\mathbf{Y}(0), \mathbf{Y}(1)]$, an $N \times 2$ array of all potential outcomes; and $\mathcal{Y}$ be the set of all possible $N \times 2$ arrays with entries of 0 or 1 so that $\mathbf{Y} \in \mathcal{Y}$. Let W_i be the treatment indicator for unit $i = 1, \ldots, N$, with

$$W_i = \begin{cases} 1 & \text{if unit } i \text{ received active treatment,} \\ 0 & \text{otherwise,} \end{cases}$$

and let $\mathbf{W}^\top = (W_1, ..., W_N)^\top$. Writing the potential outcomes as a function of only the unit indicator of treatment requires the stable unit treatment value sssumption (SUTVA Rubin, 1980), which is assumed throughout.

Assumption 1 *SUTVA: There is no interference among units or hidden varieties of treatments.*

The estimand of interest is the average causal effect,

$$\tau = \overline{Y}(1) - \overline{Y}(0),$$

where

$$\overline{Y}(w) = \frac{1}{N}\sum_i Y_i(w)$$

for $w = 0, 1$. The estimand τ is a *finite population* estimand defined in terms of potentially observable data.

We consider a completely randomized experiment where M units are randomly assigned to treatment and the other $N - M$ units are assigned to control. After assignment, we denote $y_{\text{obs},i}$ to be the observed potential outcome for unit i, that is

$$y_{\text{obs},i} = W_i Y_i(1) + (1 - W_i) Y_i(0).$$

Let $\mathbf{y}_{\text{obs}}^\top = (y_{\text{obs},1}, \ldots, y_{\text{obs},N})^\top$.

We define additional "aggregated" notation. We call $\mathbf{N} = (N^{00}, N^{01}, N^{10}, N^{11})$ the *aggregated potential outcomes* where

$$N^{yy'} = \sum_i \mathbf{1}\{Y_i(1) = y, Y_i(0) = y'\} \quad \text{for } y, y' \in \{0, 1\}.$$

The four values of the aggregated potential outcomes correspond to the number of units in the four principal strata corresponding to the "immune", "protected", "harmed", and "doomed" stratum respectively (Hudgens and Halloran, 2006; Frangakis and Rubin, 2002). For example, N^{01} is the number of units for whom treatment

protects against disease. We have $N = N^{00} + N^{01} + N^{10} + N^{11}$ and $\tau = \frac{N^{10} - N^{01}}{N}$.

Similarly, we call $\mathbf{M} = (M^{00}, M^{01}, M^{10}, M^{11})$ the *treated unit aggregated potential outcomes* where

$$M^{yy'} = \sum_i W_i \mathbf{1}\{Y_i(1) = y, Y_i(0) = y'\} \quad \text{for } y, y' \in \{0, 1\}.$$

The treated unit aggregated potential outcomes counts the number of units assigned to treatment within the four principal strata. Rosenbaum (2001) defines the effect attributable to treatment as M^{01}.

Lastly, call $\mathbf{y} = (y_\text{c}, y_\text{t})$ the *aggregated observed data* where

$$y_\text{c} = \sum_i (1 - W_i) y_{\text{obs},i} \quad \text{and} \quad y_\text{t} = \sum_i W_i y_{\text{obs},i}.$$

Our goal is to perform inference for τ using only the random assignment of treatments. We first review randomization-based inference under Fisher's sharp null (FSN).

2.2.2 A simple sharp null

Although only one potential outcome can ever be observed for any unit, a *sharp* null hypothesis, by definition, yields all potential outcomes for all observed units when combined with the observed data. The appeal of a sharp null hypothesis is that the randomization distribution of any test statistic can be determined by considering all possible randomizations.

In terms of potential outcomes, FSN of absolutely no treatment effect is expressed

Table 2.1: Observed counts of non-diseased and diseased units by treatment

	Non-diseased	Diseased	
Control	$N - M - y_{\mathrm{c}}$	y_{c}	$N - M$
Treated	$M - y_{\mathrm{t}}$	y_{t}	M
	$N - y_{\mathrm{c}} - y_{\mathrm{t}}$	$y_{\mathrm{c}} + y_{\mathrm{t}}$	N

as $Y_i(0) = Y_i(1)$ for $i = 1, \ldots, N$. Clearly, FSN is sharp. Table 2.1 characterizes the outcomes under FSN. For finite population inference, there is no debate over conditioning on margin totals in this table: row totals are fixed by a completely randomized design and column totals are fixed by virtue of FSN. Fisher's randomization test follows by noting that the (Treated, Diseased) cell in Table 2.1 has a hypergeometric$(y_{\mathrm{c}} + y_{\mathrm{t}},\ N - y_{\mathrm{c}} - y_{\mathrm{t}},\ M)$ distribution under repeated randomizations of the same units and FSN. Randomization based inference follows from considering repeated randomizations of the same set of units who have fixed potential outcomes, which are assumed known by hypothesis.

The Bayesian perspective provides another interpretation of Fisher's randomization test, with conditioning on the data following naturally. Rubin et al. (1984) formulated this test as a posterior predictive check (PPC): it is *a posteriori*, being conditional on the data, and predictive, repeating the *same* experiment with the *same* units. A PPC calculates the following probability

$$p_{\mathrm{pp}} = \Pr\big(T(\mathbf{y}_{\mathrm{obs}}^{\mathrm{pp}}, \mathbf{W}^{\mathrm{pp}}) \geq T(\mathbf{y}_{\mathrm{obs}}, \mathbf{W}) \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big), \tag{2.1}$$

where $\mathbf{y}_{\mathrm{obs}}^{\mathrm{pp}}$ and $\mathbf{W}^{\mathrm{pp}}$ are the data for the predictive experiment and T is a test statistic, for which large values correspond to extremes. Under FSN, the PPC is

equivalent to Fisher's randomization test, but the former generally provides a more flexible framework.

Rejecting FSN leads to the conclusion that there is some unit i for which $Y_i(0) \neq Y_i(1)$. Such a conclusion is not particularly interesting and leads us to consider more scientifically relevant hypotheses.

2.2.3 A Simple Dull Null Hypothesis

We first illustrate with an example of what goes wrong with a hypothesis for a specific case of $\tau = \tau_0$. Consider the hypothesis that FSN holds for all units except one, for whom the treatment protects against disease (i.e., there exists one unit i' such that $Y_{i'}(0) > Y_{i'}(1)$ and $Y_i(0) = Y_i(1)$ for $i \neq i'$). Under this hypothesis, $\tau = -\frac{1}{N}$. For units assigned to control, observing $Y_i(0) = 0$ implies $Y_i(1) = 0$, but if $Y_i(0) = 1$, then $Y_i(1) = 0$ or $Y_i(1) = 1$. Similarly, for units assigned to treatment, observing $Y_i(1) = 1$ implies $Y_i(0) = 1$, but if $Y_i(1) = 0$, then $Y_i(0) = 0$ or $Y_i(0) = 1$. We know both potential outcomes for the y_t diseased units assigned to treatment and $N - M - y_\mathrm{c}$ non-diseased units assigned to control, but we cannot determine which one of the $M - y_\mathrm{t}$ non-diseased units assigned to treatment or y_c diseased units assigned to control is the protected unit if $M - y_\mathrm{t} + y_\mathrm{c} > 1$. The simple hypothesis that treatment prevents disease for exactly one unit leads to $M - y_\mathrm{t} + y_\mathrm{c}$ arrays of possible potential outcomes.

Without a sharp hypothesis, we cannot know the randomization distribution of all statistics. Additional information can make this hypothesis sharp: if the identity of the protected unit is assumed to be known, then the remaining potential outcomes

are known. This unknown information is a nuisance because it is a fixed, unknown value that is not of primary interest. It is also missing data because it is defined in terms of potential outcomes.

2.2.4 A General Dull Null Hypothesis

As our estimand of interest is τ, testing a null hypothesis that specifies $\tau = \tau_0$ seems natural. However, this hypothesis does not uniquely determine the contrast of interest: $\overline{Y}(0) - \overline{Y}(1)$, or equivalently, $N^{10} - N^{01}$. For example, FSN implies $\tau = 0$, but a hypothesis[1] of $\tau = 0$ does not imply FSN, which is equivalent to setting *both* N^{01} and N^{10} to zero. In the spirit of FSN, we specify a hypothesis of $N^{01} = N_0^{01}$ and $N^{10} = N_0^{10}$ for $N_0^{01}, N_0^{10} \in \{0, \ldots, N\}$ and $N_0^{01} + N_0^{10} \leq N$ because together N_0^{01} and N_0^{10} determine τ_0.

As in the previous example, the hypothesis that $N^{01} = N_0^{01}$ and $N^{10} = N_0^{10}$ is not sharp in general. After observing the data, we know that the N_0^{01} units for whom treatment protects against disease are composed of some combination of the $M - y_\mathrm{t}$ non-diseased treated units or the y_c diseased control units. Also, the N_0^{10} units whom treatment harms (i.e., causes disease) are composed of some combination of y_t diseased treated units and $N - M - y_\mathrm{c}$ non-diseased control units. *A posteriori* there are

$$\binom{M - y_\mathrm{t} + y_\mathrm{c}}{N_0^{01}} \binom{y_\mathrm{t} + N - M - y_\mathrm{c}}{N_0^{10}} \tag{2.2}$$

possible arrays of potential outcomes. Again, without knowing all potential outcomes, we cannot know the randomization distribution of all test statistics under our

[1]The hypothesis of $\tau = 0$ is often referred to as Neyman's null hypothesis (Welch, 1937).

hypothesis.

Henceforth, we call the hypothesis that $N^{01} = N_0^{01}$ and $N^{10} = N_0^{10}$ the *dull null* because it is not sharp in general (and not because we think it boring!). Although the true randomization distribution is unknown under the dull null, the interpretation of Fisher's randomization test as a PPC leads to a natural testing procedure.

2.3 Inference for a dull null hypothesis

2.3.1 PPCs as extension of Fisher's randomization test

The dullness of the null hypothesis that $N^{01} = N_0^{01}$ and $N^{10} = N_0^{10}$ prevents us from using the logic of Fisher's randomization test. The analogy to PPC still applies insofar as we only need to calculate p_{pp}, as defined in equation (2.1), in order to assess the fit of the dull null. Inverting PPCs of the dull null provides a plausible region for the average causal effect. The dullness only complicates calculations.

One calculation follows from "integrating out" the specific values of the potential outcomes:

$$p_{\text{pp}} = \sum_{\mathbf{Y} \in \mathcal{Y}_0} \Pr\big(T(\mathbf{y}_{\text{obs}}^{\text{pp}}, \mathbf{W}^{\text{pp}}) \geq T(\mathbf{y}_{\text{obs}}, \mathbf{W}) \mid \mathbf{y}_{\text{obs}}, \mathbf{W}, \mathbf{Y}\big) \text{p}\big(\mathbf{Y} \mid \mathbf{y}_{\text{obs}}, \mathbf{W}\big) \qquad (2.3)$$

where $\mathcal{Y}_0 = \{\mathbf{Y} \in \mathcal{Y} : N^{01} = N_0^{01} \text{ and } N^{10} = N_0^{10}\}$ is the set of arrays of potential outcomes that satisfy the dull null hypothesis. Meng (1994) first discussed the interesting interpretation of PPCs as integrating out nuisance parameters via a two stage procedure. In particular, the first stage, corresponding to the first term in the summation, calculates a significance level under a sharp null hypothesis given all potential

outcomes, $\mathbf{Y}$. The second stage, corresponding to the second term in the summation, weights these tests of sharp nulls according to their posterior distribution. We describe computations for these quantities in more detail.

Computing the first stage

The first stage calculation requires computing a significance level for a sharp hypothesis because the array of potential outcomes, $\mathbf{Y}$, is assumed known. The significance level can be calculate using distributional theory of the test statistic or enumeration of all random assignment vectors, or it can be approximated via Monte Carlo simulations of the random assignment vectors. Significance levels must be calculated for all possible arrays of potential outcomes, of which there are a potentially very large number as demonstrate in Equation 2.2. We provide a definition of *sufficiently sharp* partitions of the set of potential outcome arrays that alleviates the need to consider every possible array of potential outcomes.

We call a partition, Π, of the set of all potential outcomes under a null hypothesis *sufficiently sharp* for statistic S, if for every $\mathcal{P} \in \Pi$, $\mathbf{Y} \in \mathcal{P}$ and $\mathbf{Y}' \in \mathcal{P}$ implies

$$\Pr\big(S = s \mid \mathbf{Y}\big) = \Pr\big(S = s \mid \mathbf{Y}'\big)$$

for all s in the support of the statistic. We use this terminology because knowing to which set in the partition an array of potential outcomes belongs is "sharp enough" to know the randomization distribution of the specified test statistic; we do not need to know the specific array of potential outcomes. A sufficiently sharp partition simplifies computation because we only need to calculate the first stage significance level for

one array of potential outcomes in each set in the partition instead of considering all possible arrays.

We provide a few examples of sufficiently sharp partitions to illustrate how this definition can be used under the dull null hypothesis. Trivially, the partition composed of singleton sets of each element, $\Upsilon_{\mathrm{s}} = \{\{\mathbf{Y}\} : \mathbf{Y} \in \mathcal{Y}_0\}$, is sufficiently sharp for all test statistics. This partition is equivalent to considering all arrays of potential outcomes. A coarser partition is required to simplify calculations.

Next consider a test statistic that is a function of the aggregated potential outcomes, $\mathbf{y}$. For example, the statistic y_{t}, as is the case for Fisher's randomization test, or $\hat{\tau} = \frac{y_{\mathrm{t}}}{M} - \frac{y_{\mathrm{c}}}{N-M}$, an unbiased estimate of the average treatment effect. We can partition the set of potential outcomes under the dull null hypothesis by the number of doomed units, N^{11}. That is, let $\mathcal{Y}_0(N_*^{11}) = \{\mathbf{Y} \in \mathcal{Y}_0 : N^{11} = N_*^{11}\}$, the set of arrays of potential outcomes with aggregated potential outcomes $\mathbf{N}_{0*} = (N_{0*}^{00}, N_0^{01}, N_0^{10}, N_*^{11})$ where $N_{0*}^{00} = N - N_0^{01} - N_0^{10} - N_*^{11}$ and $N_*^{11} \in \{0, \ldots, N - N_0^{01} - N_0^{10}\}$. The partition $\Upsilon_0 = \{\mathcal{Y}_0(0), \mathcal{Y}_0(1), \ldots, \mathcal{Y}_0(N - N_0^{01} - N_0^{10})\}$ is sufficiently sharp for $\mathbf{y}$. The proof follows from a simple representation of the distribution of $\mathbf{y}$ because if $\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})$, then

$$y_{\mathrm{c}} = N_0^{01} - M^{01} + N_*^{11} - M^{11} \tag{2.4}$$

$$y_{\mathrm{t}} = M^{11} + M^{10}, \tag{2.5}$$

which only depends on $\mathbf{N}_{0*}$ and $\mathbf{M}$. The random variable $\mathbf{M}$ follows a multivariate hypergeometric distribution with parameters $(\mathbf{N}_{0*},\ M)$ and support $\mathcal{S}_{\mathbf{N}_{0*}}$, so the

probability mass function of $\mathbf{y}$ is

$$\mathrm{p}\big(\mathbf{y} \mid \mathbf{Y}\big) = \sum_{\mathbf{M} \in \mathcal{M}_{\mathbf{y}|\mathbf{N}_{0*}}} \mathrm{p}\big(\mathbf{M} \mid \mathbf{N}\big), \tag{2.6}$$

where $\mathcal{M}_{\mathbf{y}|\mathbf{N}_{0*}} = \{\mathbf{M} \in \mathcal{S}_{\mathbf{N}_{0*}} : y_{\mathrm{c}} = N_0^{01} - M^{01} + N_*^{11} - M^{11} \text{ and } y_{\mathrm{t}} = M^{11} + M^{10}\}$. The distribution of $\mathbf{y}$ only depends on the array of potential outcomes, $\mathbf{Y}$, through the values of the aggregated potential outcomes, $\mathbf{N}$, which are the same for all elements of $\mathcal{Y}_0(N_*^{11})$. Therefore, the partition Υ_0 is a sufficiently sharp partition for $\mathbf{y}$ under all dull null hypotheses. Using this sufficiently sharp partition reduces the number of first stage significance levels that need to be calculated to $N - N_0^{10} - N_0^{01} + 1$ instead of the number in (2.2). The partition is also sufficiently sharp for any function of $\mathbf{y}$.

Without any additional information, most test statistics in a completely randomized experiment will be a function of $\mathbf{y}$. A test statistic that is a function of the first $K < N$ units is not a function of $\mathbf{y}$ (and Υ_0 is not sufficiently sharp). If K is arbitrarily chosen, this statistic is a little odd because it ignores some of the data. However, if K is chosen to correspond to a pretreatment covariate, for example sex of the unit, then it would be reasonable to consider such statistics. Sufficiently sharp partitions exist for statistics that are functions of covariates.

Computing the second stage

The posterior of $\mathbf{Y}$—the second term after the summation in equation (2.3)—is proportional to the likelihood times the prior distribution:

$$\mathrm{p}\big(\mathbf{Y} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big) \propto \mathrm{p}\big(\mathbf{y}_{\mathrm{obs}}, \mathbf{W} \mid \mathbf{Y}\big)\ \mathrm{p}\big(\mathbf{Y}\big)$$

where all probabilities assume the dull null hypothesis. The likelihood is

$$\mathrm{p}\big(\mathbf{y}_{\mathrm{obs}}, \mathbf{W} \mid \mathbf{Y}\big) = \mathrm{p}\big(\mathbf{y}_{\mathrm{obs}} \mid \mathbf{W}, \mathbf{Y}\big)\mathrm{p}\big(\mathbf{W}\big) = \begin{cases} \binom{N}{M}^{-1} & \text{if } \mathbf{y}_{\mathrm{obs}} \text{ follows from } \mathbf{W} \text{ and } \mathbf{Y}, \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

because $\mathbf{y}_{\mathrm{obs}}$ is a deterministic function of $\mathbf{W}$ and $\mathbf{Y}$, and $\mathbf{W}$ is assigned using a completely randomized design. The observed data restricts the set of possible arrays of potential outcomes.

Choosing a prior distribution for the array of potential outcomes satisfying the dull null hypothesis is tricky because it is difficult to conceptualize the space of all such arrays. Instead, we prefer to place a prior distribution on an interpretable function of the potential outcomes, which may be lower dimensional. We suggest placing a prior distribution on the sets of a sufficiently sharp partition for a test statistic, with all elements within a set treated symmetrically. Although it seems strange to draw prior information from the choice of test statistic, we believe the choice of test statistic provides prior information in the form of possible violations of the hypothesis and

symmetries on units.

We illustrate the difficulties of choosing a prior distribution and our proposed prior distribution under the dull null hypothesis and a test statistic that is a function of $\mathbf{y}$. First consider a "non-informative" uniform prior distribution on all arrays of potential outcomes, that is $\mathrm{p}\big(\mathbf{Y}\big) \propto 1$ for $\mathbf{Y} \in \mathcal{Y}_0$. This prior distribution induces a very informative binomial$(N - N_0^{01} - N_0^{10}, \frac{1}{2})$ prior distribution on the number of doomed units, N^{11}. The quantity N^{11} is more interpretable than a specific array of potential outcomes and corresponds to a unique set in the sufficiently sharp partition for $\mathbf{y}$. We propose placing a prior distribution on the event $\{\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\}$ and equal probabilities to each element within that set (i.e., $\mathrm{p}\big(\mathbf{Y}\big) = \mathrm{p}\big(\mathbf{Y}'\big)$ if $\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})$ and $\mathbf{Y}' \in \mathcal{Y}_0(N_*^{11})$). This prior distribution reduces to placing a prior distribution on N^{11}. We recommend a uniform distribution over all plausible values of N^{11}.

Computing $p_{\mathbf{pp}}$

Considering the sufficiently sharp partition in both the first and second stage computation simplifies the calculation and interpretation of the posterior predictive p-value. Using a test statistic that is a function of $\mathbf{y}$ and a prior distribution on the sufficiently sharp partition Υ_0, we can show

$$\Pr\big(T(\mathbf{y}^{\mathrm{pp}}) \geq T(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big) = \Pr\big(T(\mathbf{y}^{\mathrm{pp}}) \geq T(\mathbf{y}) \mid \mathbf{y}\big), \tag{2.8}$$

using a counting argument (details in Appendix B).

The sufficiently sharp partition is inversely related to the idea of data reduction via sufficient statistics. In the latter case, a parametric model induces a sufficient statistic

that "contains the whole of the information about the parameter". The sufficiently sharp partition applies for a statistic first and then partitions the space of arrays of potential outcomes—or "parameters"—into sets that contain all the information for the test statistic.

The extension of Fisher's randomization test to PPCs maintains the logic of using randomization as the inferential tool and allows us to assess the fit of a dull null hypothesis. Lack of fit can be attributed to violations of the dull null hypothesis, the likelihood, or the prior distribution. We view the likelihood as assumption free because it is derived from the randomization of a finite population. If the PPCs are not too sensitive to the prior distribution (we examine prior sensitivity in Section 2.4.2), inverting PPCs for different values of N_0^{01} and N_0^{10} leads to a plausible region for the average causal effect.

2.3.2 Other methods of handling nuisances

Several methods have been proposed for the well known problem of nuisance parameters (see Basu, 1977). We cover some approaches here, but do not provide an exhaustive review.

Instead of averaging over a set of arrays of potential outcomes, we could maximize over this set. Define the *p-value* as the largest significance level when considering all possible values of the nuisance (Casella and Berger, 2001, p. 397). That is,

$$p_{\max} = \max_{\mathbf{Y} \in \mathcal{Y}_0} \Pr_{\mathbf{Y}}\big(T(\mathbf{y}_{\text{obs}}^{\text{rep}}, \mathbf{W}^{\text{rep}}) \geq T(\mathbf{y}_{\text{obs}}, \mathbf{W})\big),$$

where the probability refers to repeated randomizations for a fixed potential outcome

array. The p-value guarantees the type 1 error rate is no larger than a prescribed level but is often criticized for being too conservative because it considers values of the nuisance that are not relevant to the observed data.

To alleviate the cautiousness of the p-value, the *plug-in* method assumes the nuisance parameter is fixed at an estimate. As discussed in the previous section, there is no evidence for or against any specific array of potential outcomes, so estimating a specific value of the nuisance is difficult. However, the value N^{11} can be estimated, so a possible plug-in procedure might maximize over all potential outcomes with N^{11} fixed at an estimate $\widehat{N}^{11}$, that is

$$p_{\text{plug}} = \max_{\mathbf{Y} \in \mathcal{Y}_0(\widehat{N}^{11})} \Pr_{\mathbf{Y}}\big(T(\mathbf{y}_{\text{obs}}^{\text{rep}}, \mathbf{W}^{\text{rep}}) \geq T(\mathbf{y}_{\text{obs}}, \mathbf{W})\big).$$

Plug-in methods test sets of potential outcomes that are more relevant to the observed data, but they do not account for the uncertainty of the estimate of the nuisance.

Perhaps the most appealing solution to handling nuisances is to account for the additional uncertainty in the reference distribution of a test statistic using a pivot (e.g., Student's t-test and Pearson's χ^2 test of independence). We have not been able to find such a test statistic and its corresponding reference distribution for this problem.

No discussion of interval estimation for average causal effects would be complete without considering Jerzy Neyman's important contributions (Neyman et al., 1990; Neyman, 1934). Neyman's approach can be summarized as follows (Imbens and Rubin, 2014, chapter 6): first find an unbiased estimator of the estimand, calculate the variance of the estimator under repeated randomizations, find an unbiased or

positively biased estimator of the variance, and appeal to finite population central limit theorem to get confidence interval estimates. Neyman et al. (1990) showed the variance of the unbiased estimator $\hat{\tau}$ cannot be unbiasedly estimated. A positively biased estimator corresponds to (but does not require) an assumption of constant additive treatment effects.

Models of binary outcomes can provide useful interval estimates of finite population estimands. The common confidence interval for the difference of two proportions is procedurally similar to Neyman's approach, but assumes a super population binomial model on the aggregated observed outcomes. Exact inference for the same super population model avoids nuisance parameters by changing the estimand of interest to the odds ratio (Cox and Snell, 1989, Section 2.3). The phenomenological Bayesian approach directly imputes missing potential outcomes with a model and provides interval estimates from the posterior distribution of the estimand (Rubin, 1978a,b). With the exception of phenomenological Bayes, model-based approaches do not explicitly discuss the role of randomization.

2.4 Examples with 42 units

We illustrate our procedure on some toy examples for completely randomized experiments, in particular when $N = 42$. First we demonstrate an analysis of a real dataset and then consider repeated sampling properties of PPCs. We focus on non-asymptotic approaches and exact calculations for such a small sample size.

2.4.1 One real toy dataset

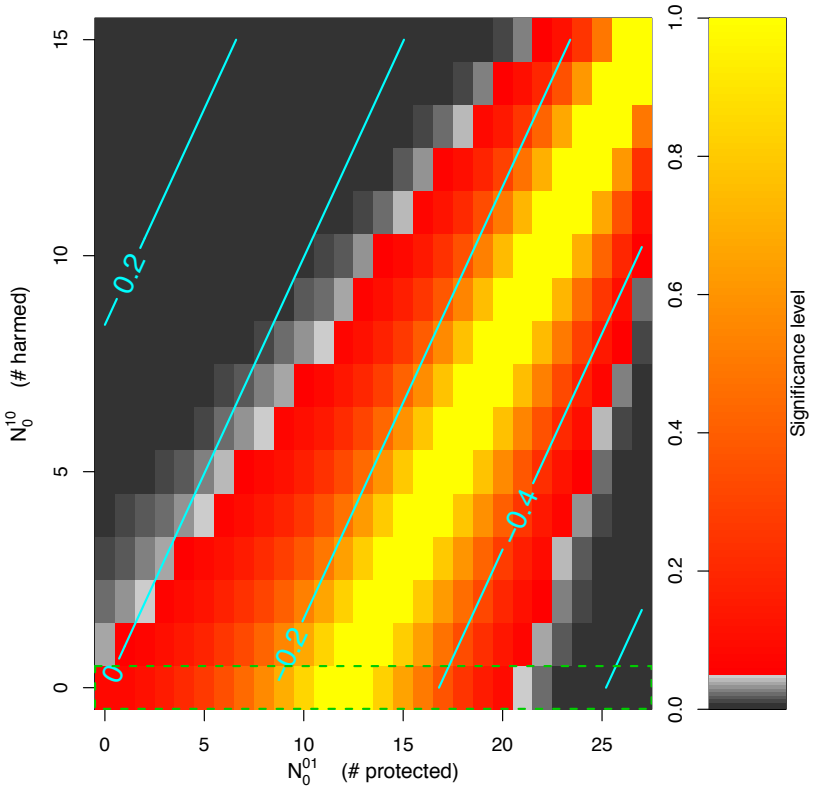

Figure 2.1: Significance levels for all possible dull null hypotheses for the Freireich dataset. Light blue contours highlight values of τ. The non-grey area corresponds to inverting p_{pp} at the $\alpha = 0.05$ level. The region highlighted in green dashes corresponds to the assumption of monotonicity.

We illustrate our approach with an analysis of data from a clinical trial of palliative therapy for acute leukemia from Freireich et al. (1963). Gehan (1965) and later Cox (1972) canonized this dataset with demonstrations of survival analysis with right censoring. Because we are concerned with binary outcomes, we consider the outcome as either in remission (i.e., non-diseased) or not in remission (i.e., diseased) at six weeks from the start of remission. The cutoff at six weeks is made to obviate the

problem of right censoring—remission status at six weeks is observed for all units. We assume a completely randomized design was used when in fact the trial used a paired and sequential design.

This study involved $N = 42$ units who recently went into remission. Half of the units were assigned treatment ($M = 21$) and half were assigned placebo. After six weeks, $y_{\text{t}} = 3$ units ended remission in the treated group and $y_{\text{c}} = 9$ units ended remission in the control group.

We invert PPCs of the dull null hypothesis to get a plausible region for the average causal effect. For a test statistic, we use the difference of the sample means: $T(\mathbf{y}_{\text{obs}}, \mathbf{W}) = \hat{\tau}$. Both large and small values of this statistic are extreme, so we consider significance levels of the form $2 \cdot \min(p^{\text{l}}_{\text{pp}}, p^{\text{g}}_{\text{pp}})$, where

$$p^{\text{l}}_{\text{pp}} = \Pr\big(T(\mathbf{y}^{\text{pp}}_{\text{obs}}, \mathbf{W}^{\text{pp}}) \leq T(\mathbf{y}_{\text{obs}}, \mathbf{W}) \mid \mathbf{y}_{\text{obs}}, \mathbf{W}\big)$$
$$p^{\text{g}}_{\text{pp}} = \Pr\big(T(\mathbf{y}^{\text{pp}}_{\text{obs}}, \mathbf{W}^{\text{pp}}) \geq T(\mathbf{y}_{\text{obs}}, \mathbf{W}) \mid \mathbf{y}_{\text{obs}}, \mathbf{W}\big).$$

We place a uniform prior distribution on the sets of Υ_0, a sufficiently sharp partition for $\hat{\tau}$.

For these data, the range of plausible values for N^{01} is 0 to 27 and for N^{10} is 0 to 15. We calculate p_{pp} for every dull null hypothesis comprised of a combination of these two values. The results are plotted in Figure 2.1, in which non-grey areas correspond to plausible dull null hypotheses at the 0.05 level. From this figure, it is clear that no one dimensional interval can summarize the plausible region for τ. The widest possible interval for τ corresponds to (-0.48, 0).

2.4.2 All possible toy datasets

So far, we have refrained from describing intervals from inverted PPCs as having a specified confidence level. Such discretion is required because PPCs do not generally maintain type 1 error (Rubin et al., 1984; Meng, 1994; Gelman et al., 1996; Bayarri and Berger, 2000; Robins et al., 2000), and thus intervals from inverted PPCs will not guarantee a nominal coverage. In order to assess the repeated sampling properties of these intervals, we calculate the type 1 error rate of these tests, which corresponds to one minus the coverage.

We focus on all possible aggregated potential outcomes. Results for specific arrays of potential outcomes follows by symmetry. For $N = 42$, the set of all aggregated potential outcomes is $\{\mathbf{N} \in \mathbb{N}^4 : N^{00} + N^{01} + N^{10} + N^{11} = 42\}$, which corresponds to a 3-simplex over integer values. Figure 2.2 illustrates this set, which contains 14,190 elements. We examine two cases. First, we assess the fit of the model which assumes the dull null hypothesis. Then we consider one overall test for a specified value of the average causal effect.

Testing the true dull null

For all possible aggregated potential outcomes, we evaluate the type 1 error rate when the dull null is true. That is, for $\mathbf{Y}_0$ with aggregated potential outcomes $\mathbf{N} = (N_0^{00}, N_0^{01}, N_0^{10}, N_0^{11})$, we calculate

$$\text{Type 1 Error Rate} = \Pr_{\mathbf{Y}_0}\big(p \leq \alpha\big), \tag{2.9}$$

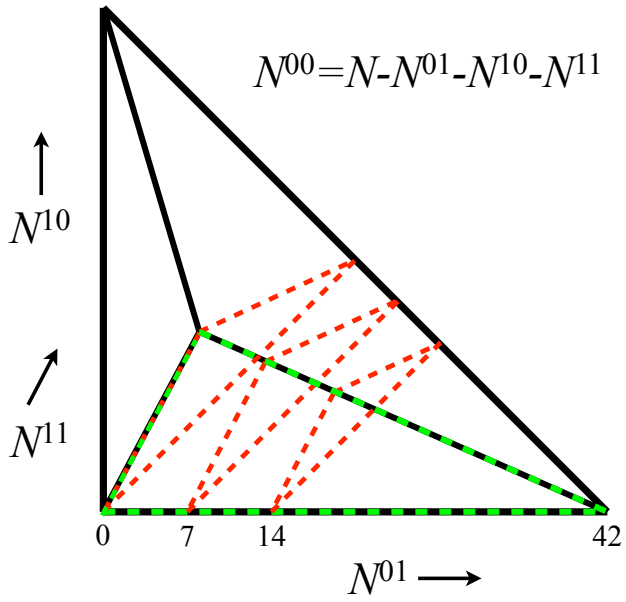

Figure 2.2: Graphical representation of the set of all aggregated potential outcomes for $N = 42$. The plane highlighted in green dashes is the subset of aggregated potential outcomes where monotonicity holds. The planes highlighted in red dashes are subsets in which τ equals 0, -1/6, and -1/3 from left to right respectively.

where p is a significance level of a test that assumes the true dull null hypothesis of $N^{01} = N_0^{01}$ and $N^{10} = N_0^{10}$ (and ignores N_0^{11} and N_0^{00}) and the probability is taken over randomizations under the true array of potential outcomes. We use the same test statistic, significance level, and prior distribution as the previous section. The number of treated units is varied: $M = 21$, 28, 32, and 35 corresponding to allocations of treatment to control of 1:1, 2:1, 3:1, and 5:1 respectively. We examine type 1 error rates for $\alpha = 0.05$. For comparison, we calculate type 1 error rates for the plug-in method for two estimators of N^{11}: $p_{\text{plug(MAP)}}$ plugs in the maximum a posteriori (MAP) estimator and $p_{\text{plug(MoM)}}$ plugs in a method of moments (MoM) estimator[2].

Table 2.2: Summaries of type 1 error rates for testing the true dull null hypothesis. All possible aggregated potential outcomes are considered for $N = 42$.

Procedure	$M = 21$	$M = 28$	$M = 32$	$M = 35$
	Proportion with type 1 error > 0.05			
p_{pp}	0.005	0.001	0.029	0.038
$p_{\text{plug(MAP)}}$	0	< 0.001	0.003	0.005
$p_{\text{plug(MoM)}}$	0.043	0.021	0.143	0.071
	Maximum type 1 error			
p_{pp}	0.052	0.052	0.059	0.061
$p_{\text{plug(MAP)}}$	0.05	0.052	0.058	0.065
$p_{\text{plug(MoM)}}$	0.114	0.08	0.141	0.153

Summaries of type 1 error rates are presented in Table 2.2. We summarize the results for $M = 21$. For PPC, 0.5% of 14,190 aggregated potential outcomes have a type one error greater than the nominal 0.05 level. The largest type 1 error rate is 0.052. The plug-in method works well when using the MAP estimator of the nuisance

[2]We round the unbiased MoM estimator, $\widehat{N}^{11}_{\text{MoM}} = \frac{N}{M} y_{\text{t}} - N_0^{01}$, to the nearest plausible integer.

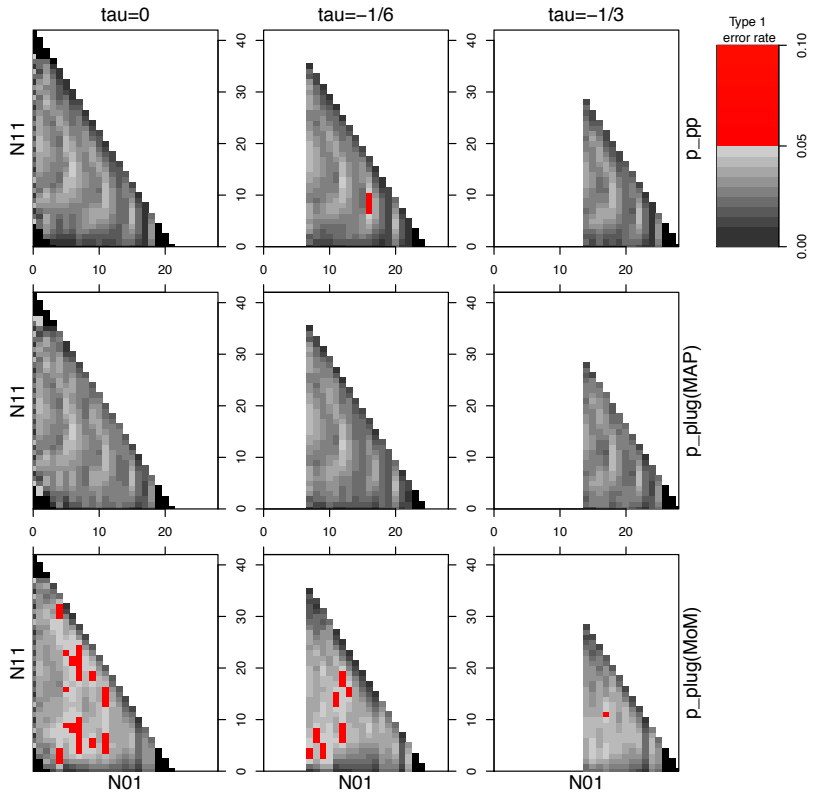

Figure 2.3: Type 1 error rates for p_{pp}, $p_{\text{plug(MAP)}}$, and $p_{\text{plug(MoM)}}$ (from top to bottom respectively) for aggregated potential outcomes when $N = 42$, $M = 21$, and τ=0, -1/6, and -1/3 (from left to right respectively).

parameter, but not as well with the naive MoM estimator. For the plug-in method using the MoM estimator, 4.3% of all aggregated potential outcomes have type 1 error greater than 0.05 (with a largest rate of 0.114), whereas using the MAP has no type 1 error rates greater than 0.05. As M gets closer to N, the procedures perform less well, but even the maximum type 1 error rates for p_{pp} and $p_{\text{plug(MAP)}}$ are still close to the nominal level. Type 1 error rates for $M = 21$ are plotted in Figure 2.3 for aggregated potential outcomes with $\tau = 0$, -1/6, and -1/3. The triangles from Figure 2.3 correspond to the dashed red triangles in Figure 2.2.

The only difference between the assumptions underlying the PPCs and the true set of potential outcomes is our prior distribution on the potential outcomes. The repeated sampling properties of the PPCs are not too sensitive to our choice of prior because the type 1 error rates are near the nominal level. Therefore, small significance levels can be attributed to either a lack of fit of the dull null hypothesis or a rare event occurring.

A single test for the true average causal effect

Enumerating tests over all possible values of N^{01} and N^{10} can be tedious. To alleviate this arduous task, for a specific value of the average causal effect we propose using one hypothesis that assumes monotonicity on the potential outcomes.

Assumption 2 *Monotonicity: The treatment cannot harm units. That is,* $Y_i(0) \geq Y_i(1)$ *for all* $i = 1, \ldots, N$.

For a specific $\tau_0 \leq 0$, this procedure tests the dull null $N^{01} = -N\tau_0$ and $N^{10} = 0$. Of course, the monotonicity assumption can also assume treatment cannot protect

units, which corresponds to a nonnegative average causal effect.

We evaluate type one error properties of this procedure for tests that assume the true average causal effect. That is, for $\mathbf{Y}_0$ with aggregated potential outcomes $\mathbf{N} = (N_0^{00}, N_0^{01}, N_0^{10}, N_0^{11})$, we calculate the probability in equation (2.9), but now p is the significance level from a test that assumes the dull null of $N^{01} = N_0^{01} - N_0^{10}$ and $N^{10} = 0$. Type 1 error rates using this testing procedure are presented in Table 2.3. This procedure is valid for almost all arrays of potential outcomes, even when monotonicity does not hold. By valid, we mean the type one error rate is at most the nominal level.

Table 2.3: Summaries of type 1 error rates for testing $N^{01} = N\tau_0$ and $N^{10} = 0$. All possible aggregated potential outcomes are considered for $N = 42$.

Procedure	$M = 21$	$M = 28$	$M = 32$	$M = 35$
	Proportion with type 1 error > 0.05			
p_{pp}	0	< 0.001	0.002	0.001
$p_{\text{plug(MAP)}}$	0	0	0.001	< 0.001
$p_{\text{plug(MoM)}}$	0.007	0.003	0.036	0.003
	Maximum type 1 error			
p_{pp}	0.05	0.051	0.056	0.058
$p_{\text{plug(MAP)}}$	0.05	0.049	0.054	0.052
$p_{\text{plug(MoM)}}$	0.114	0.072	0.087	0.068

This procedure being valid is not surprising. In Section 2.4.1, the widest plausible interval for τ corresponds to tests under the assumption of monotonicity (see the region highlighted in green in Figure 2.1). Furthermore, the monotonicity assumption is analogous to (and weaker than) the assumption of constant additive treatment effects for continuous outcomes. The latter assumption corresponds to valid confidence intervals from Neyman's perspective, even when additivity does not hold.

We have shown for the case of $N = 42$ that inverting hypothesis tests using PPCs or the plug-in method using the MAP estimator will give confidence intervals with approximate nominal coverage. Although rigorous theory would be ideal, our exhaustive study for a typical value of N shows our procedure generally provides valid confidence intervals.

2.5 Discussion

Although we have focused on the simple case of a completely randomized experiment with the estimand of average causal effects, using PPCs to test non-sharp nulls of causal effects applies generally for estimands involving binary outcomes. The general approach defines a hypothesis for a specific estimand of interest. The PPC averages sharp hypothesis tests for every array of potential outcomes satisfying the dull null hypothesis with respect to the posterior distribution. The posterior distribution is proportional to a prior distribution on the potential outcomes times an assumption free likelihood that is induced by randomization.

With this general framework, we can address other experimental designs. Extending inference to randomized block and paired designs is simple. Another randomization scheme to consider is the group sequential design used by data and safety monitoring boards (Pocock, 1977; O'Brien and Fleming, 1979).

In addition to more designs, we can address other causal estimands of interest. Although unit-level relative risks (and odds ratios) are ill-defined, we can consider

the following finite population relative risk:

$$\tau_{\text{rr}} = \frac{\sum_i Y_i(0)}{\sum_i Y_i(1)}.$$

Other estimands might pertain to a specific subset of units as is the case when estimands are defined with respect to a principal stratum (Frangakis and Rubin, 2002). Examples of such estimands are the complier average causal effect, which arises when units do not perfectly comply with the assigned treatment (Sommer and Zeger, 1991; Angrist et al., 1996); the survivor average causal effect, which arises when outcomes are "censored by death" (Rubin, 2006; Zhang and Rubin, 2003); and the effect of vaccine on disease severity, which is only well-defined for the doomed principal stratum (Hudgens and Halloran, 2006).

One limitation to this approach is that it may be difficult to incorporate continuous covariates or outcomes because any one value will not have enough units to create a group. However, discretizing continuous values can offer robustness to model assumptions (Cangul et al., 2009). Moreover, computation of PPCs is more intensive than standard asymptotic approximations or super population methods. Lastly, placing a prior distribution on the array of potential outcomes is difficult. Using sufficiently sharp partitions for a given test statistic can both simplify computation and motivate possible prior distributions.

Our focus on a finite population may also seem a limitation because any experiment has the goal of generalizing to some larger population. Although such transportability of results is important, we view our inference as derived from only the physical randomization controlled by the experimenter. Without a clearly defined

super population and a random sampling scheme of that population, assumptions about the representativeness of the experimental units must be made. We eschew such assumptions here and opt for principled analysis of the finite population.

These complications do not arise from trying to make a simple problem more difficult, but rather originate from the reasonable approach of considering a finite population and physical randomization as our inferential tool. Interpreting Fisher's randomization test as PPCs leads to a natural test that integrates out nuisance unknowns. By inverting these hypothesis tests, we create approximate confidence intervals for estimands involving binary outcomes.

Chapter 3

Exposure efficacy and interference in prophylactic treatments of HIV

3.1 Introduction

A treatment that reduces susceptibility to HIV transmission could have a massive impact on the HIV/AIDS epidemic. Development of an effective vaccine for HIV has been elusive (Walker and Burton, 2008; Hammer et al., 2013; Buchbinder et al., 2008). In lieu of a vaccine, a dose of antiretroviral agents has been proposed as a pre-exposure prophylactic (PrEP) treatment that would be used by uninfected persons before (or soon after) sexual intercourse with potentially HIV infected partners (Cohen and Baden, 2012). Some clinical trials have provided evidence for PrEP treatments effectively preventing HIV infections (Thigpen et al., 2012; Karim et al., 2010; Grant et al., 2010; Baeten et al., 2012), but other trials have been stopped early because effectiveness could not be established (Cohen and Baden, 2012; Van Damme et al.,

2012). Further research is needed to examine the potential effects of PrEP treatments as well as new vaccines. We discuss the difficulties in assessing the effect of a prophylactic treatment of HIV from the perspective of the Rubin Causal Model (RCM, Holland, 1986), which extends the causal model of Neyman et al. (1990) beyond completely randomized experiments. Assessing efficacy of prophylactic treatments of HIV fits into the more general framework of assessing efficacy of a preventative treatment of an infectious disease (for a full length text on the subject, see Halloran et al., 2010). We focus on two issues in such studies: exposure to disease and interference between units.

The first issue is concerned with whether or not a person is exposed to the disease. Being exposed to the disease is of obvious importance for assessing the efficacy of the disease. Exposure occurs after treatment is assigned, and thus it is a well defined outcome. Differences of exposure under different treatments imply an effect of assignment to treatment on exposure. The potential for such effects on exposure are well known and precautions are taken to ensure some intuitive notion of comparable exposure under both treatments. Our contribution follows the RCM by conceptualizing the potential outcomes of exposure for a given subject under both the assignment to the experimental and control treatment as well as the treatment assignment to the entire population. With the additional notation for exposure, we show that identical exposure cannot occur in general for an effective treatment.

The second issue with trials of infectious disease is that the treatment of one unit can interfere with the outcome of infection for another unit. We link the ideas of exposure to an infectious disease and interference between units: a prophylactic

treatment has the potential to prevent disease for some people and thus prevents these people from exposing others to the disease. The reason for interference for the outcome of infection is due to interference on the amount of exposure for another unit. Hudgens and Halloran (HH, 2008) describe the general problem of inference between units using the RCM to define estimands that describe the direct, indirect, total, and overall effects of the treatment with respect to different intervention programs. The indirect effect of HH can be interpreted as the quantifying interference in the randomized experiment. Our contribution is a simple threshold model for disease transmission that considers the number of exposures before infection. The threshold model can be related to the probabilistic binomial model for disease transmission, but has the advantage of producing fixed outcomes for each combination of treatment assignments, even in the presence of interference. Fixed outcomes is a requirement for the estimands within the RCM. We perform a simulation study of a population and quantify the amount of interference in a simulated population using the estimands of HH.

We focus on the case study of a prophylactic treatment of HIV as a motivating example because the two issues of exposure and interference have physical interpretations. Exposure to HIV can be well defined as a distinct event occurring in continuous time. In particular, we consider exposure to HIV for an uninfected unit as unprotected sexual intercourse with an HIV infected unit. The clear description of exposure also allows us to consider how units can potentially interfere with each other. In particular, interference between units can be described via a network of sexual partners.

In the next section, we provide intuition behind the complications that arise in trials for preventive treatments of infectious disease. Section 3.3 describes general potential outcomes notation with which we can define more clearly the post-treatment outcome of exposure. Section 3.4 describes a simple example of a couples study design in which the complications of interference are avoided in order to focus on the specific issue of exposure. Section 3.5 describes another more general example of a closed population design in which interference does occur. We introduce the threshold model and demonstrate with a simulation study how the concepts of exposure and interference are related.

3.2 Challenges

The challenges of causal inference of infectious diseases have been addressed in many settings. We introduce some of the conceptual challenges that arise with examples and address how we approach these issues for a trial of the efficacy of a prophylactic treatment of HIV. More formal definitions and examples using the notation of the RCM will be considered in later sections.

3.2.1 Defining Exposure

For any given infectious disease, "being exposed" to the disease can be hard to define. Consider influenza as an example. What does it mean to be in contact with, or exposed to the flu? Does one need to be in close contact with an infected person? Or is everyone in a room exposed if there is one infected person in the room?

We follow Rhodes et al. (1996) and define an exposure for a susceptible person as

a type of contact with another person (or persons) who is (or are) infectious. More specifically, a susceptible person is someone who can potentially contract the disease (i.e., not already infected) and an infectious person is someone who can potentially infect another person. A contact is some interaction in which the disease could be transmitted between the infectious person and susceptible person. Different contacts necessarily define different types of exposures.

For example, Longini et al. (1988) differentiate two types of exposures: within a household and within a community. For household exposure, contact is defined by living in close proximity to another person in a household. An household is infectious if one person in the household is infectious; that is all of the other members of that household are considered exposed if one person in the household is infectious. In contrast, community exposure represents a person's everyday interaction within a community. Everyone in the same community receives the same amount of exposure at the community level, so the community is always infectious.

Definitions of exposure to HIV are more clear. HIV is primarily transmitted through unprotected sexual intercourse or sharing intravenous needles, thus the contacts can be defined as unprotected sex or sharing needles. Other types of contacts can be defined or further distinguished (e.g., vaginal sex, oral sex, etc.). Furthermore, an HIV infected person is infectious.

For this discussion, we consider one definition of exposure to HIV as unprotected sex with an HIV infected individual. We focus on a population of men who have sex with men, which is a common target population for trials of PrEP treatments. Each contact is composed of two individuals in the population and a unique time at which

the sexual encounter occurred. Additionally, we assume that all HIV infected persons are infectious and that if infection occurs for a susceptible unit, it occurs at a distinct exposure. These notions are made more concrete in Section 3.3.1.

3.2.2 Documenting Exposure

Even if exposure can be well defined for a given contact, actually documenting whether or not a person is exposed is not trivial. In the example of household exposure from Longini et al. (1988), if contact is defined as living in the same household as an infected person, then we must document both household membership and the infectiousness of every member of the household. For a given unit, both the data on contacts and infection status of contacted persons must be documented in order to assess exposure.

For HIV, this information amounts to knowing the number of sex acts a unit participates in and whether or not the unit's sexual partners are HIV infected. A unit will most likely know the former information, but often not the latter. Furthermore, collecting information on the HIV status of sexual partners for every unit in the study would be expensive and unreasonable. We proceed under the assumption that information on the number of sex acts can be documented (i.e., the contact process), but the infection status of sexual partners cannot be directly documented.

Whether or not an exposure can be documented is a practical concern, in particular for estimation of causal effects. Rhodes et al. (1996) discuss different models that condition on exposure information, which we interpret as documenting exposure information under the assigned treatment. In contrast, we focus on the description of

exposure under different treatment assignments and possible differences across treatment assignments.

3.2.3 Identical Exposure

We are focused on treatments that attempt to reduce susceptibility to a disease, that is the treatment attempts to reduce the chance of contracting the disease when exposed. In such situations, the *identical exposure condition* assumes every subject is exposed equally to the disease under assignment to the active and control treatments. The identical exposure condition ensures that the comparison of subjects assigned to treatment and subjects assigned to control have the same amount of exposure (on average). Differences in exposure across treatment groups could imply a different mechanism for reducing infections than the desired effect on susceptibility. We return to this concept in more detail in Section 3.4.2.

Historically, Greenwood and Yule (1915) first described the identical exposure condition, and we briefly contrast this condition to another requirement of Greenwood and Yule (1915): "[t]he persons must be, *in all material respects*, alike." We interpret the distinction between these two conditions on pre-treatment covariates and post-treatment variables (i.e., exposure). The examples provided in Greenwood and Yule (1915) suggest the distinction is made for practical reasons stemming from observational studies (e.g., comparing inoculated units to uninoculated units across different years) instead of distinguishing between pre-treatment and post-treatment quantities in a randomized experiment, which had not been formally defined at the time. Regardless of the intentions of Greenwood and Yule (1915), the notion of iden-

tical exposure is a valid request in a clinical trial of a prophylactic treatment. Next, we discuss some violations of identical exposure.

3.2.4 Effects on Exposure

Exposure occurs after treatment has been assigned and thus may be affected by the treatment. Exposure effects are a violation of the identical exposure condition, and it is important to consider possible effects—both intended and unintended—of the treatment on exposure. Indeed, many interventions for infectious diseases directly target reducing the number of exposures. For example, personalized counseling for adolescents has been shown to reduce high-risk sexual behaviors, or contacts (Kamb et al., 1998). The identical exposure condition is required for treatments that reduce susceptibility and thus do not directly target reductions in exposure. Halloran et al. (1994) introduce this notion of exposure efficacy and describe possible biases that can arise in simple models of disease transmission when the treatment affects exposure. We consider possible explanations for treatment effects on exposure, but do not address these issues in more detail.

One concern for developing an effective HIV prophylactic is the potential for risk compensation, or the increase of risky behavior because of a presumed decrease in susceptibility (Cassell et al., 2006). That is, a person may engage in more risky sexual practices if he or she knows they are taking a potentially effective prophylactic treatment. Using placebos and blinding subjects to the treatment attempt to protect against such effects as well as additional counseling for HIV prevention regardless of a subjects' treatment assignment. Guest et al. (2008) present results on possible risk

compensation in a PrEP trial.

A more subtle effect of treatment on exposure may be due to side effects caused by the active treatment. If the treatment causes a side effect that leads to a change in a person's sexual activity, then the treatment might (inadvertently) cause a decrease in exposure to the disease. For example, Karim et al. (2010) showed a PrEP treatment might cause diarrhea in some subjects, who may have changed their sexual behavior as a result.

There are several potential effects of the treatment on exposure. The above examples explain possible treatment effects on units' sexual behavior and thus the way that units contact each other. In contrast, we now consider an effect on exposure that changes the infectiousness of partners.

3.2.5 Interference between units

Interference between units in a trial of a prophylactic treatment of HIV can be attributed to an effect on exposure. If the treatment prevents one unit from contracting HIV, then treatment is also preventing that unit from exposing additional people in the population. The treatment reduces exposures by reducing the number of infectious people in the population. An important point about interference between units in studies of infectious disease is that it can be attributed to an effect on exposure.

For a trial of a prophylactic for HIV, units expose each other to HIV via sexual contacts. Furthermore, by considering contacts as events occurring at specific times between two units, we can conceive of exposure to HIV occurring within a network of sexual contacts. Many studies have examined sexual networks and its relation to

sexually transmitted diseases (Jones and Handcock, 2003; Helleringer and Kohler, 2007). Additional work has considered the spread of epidemics through networks (Newman, 2002; Ganesh et al., 2005; Kenah and Robins, 2007; Volz and Meyers, 2009). We distinguish our contribution as examining the network of sexual contacts as creating interference in a trial of prophylactic treatment for HIV.

In Section 3.3, we follow prior work on the RCM in the presence of interference. In particular, Sobel (2006) considered estimands and estimation in situations that violate the common no interference assumption (a.k.a., part of the stable unit treatment value assumption, Rubin, 1980). HH generalized some of the ideas of Sobel (2006) and defined estimands describing the direct, indirect, total, and overall effects of the treatment. We describe these estimands in more detail in Section 3.5.1. Rosenbaum (2007) offered a different inferential goal by describing the effect attributable to interference. Our work extends ideas introduced in Halloran and Struchiner (1995), which also considers the role of exposure in interference, by describing how indirect effects on exposure result in indirect effects on infections. This connection is made through the threshold model, introduced in Section 3.5.2, and demonstrated with a simulation in Section 3.5.3.

3.3 Neyman-Rubin Causal Model

We make this discussion more concrete by developing notation of the RCM for a simple example of an experiment that tests the effect of a prophylactic treatment of HIV. The RCM conceptualizes the "potential outcomes" under all possible treatments for each unit. We consider a trial that begins and ends at a fixed time.

3.3.1 Potential outcomes

We consider a finite population of N units. The population is divided into an experimental population and non-experimental population. The experimental population consists of $n \leq N$ units to whom the experimenter assigns either an experimental treatment or control treatment. The experimental population is a fixed subset of the population, that is, we do not consider how the experimental population is sampled (random or otherwise) from the finite population. Generalizing the results of the experimental population to a new population relies on assumptions of how "comparable" the new population is to the experimental population. The non-experimental population is composed of all units who potentially come into contact with the experimental population for the duration of the trial. The distinction between the non-experimental and experimental populations is necessary because units in the experimental population can interact with units outside the experimental population and the outcomes of the experimental population may depend on the outcomes of the non-experimental population. For convenience, we index the experimental population with $i = 1, \ldots, n$ and the remaining non-experimental population with $i = n + 1, \ldots, N$.

Each unit in the experimental population receives one of two possible treatments. Let Z_i denote the treatment that unit $i = 1, \ldots, n$ receives where 1 corresponds to the active treatment and 0 corresponds to the control treatment. Units in the non-experimental population cannot receive either treatment and thus we let $Z_i = *$ for $i = n + 1, \ldots, N$. We let $\mathbf{Z} = (Z_1, \ldots, Z_N)$ be the treatment assignment vector. Let

$\mathcal{Z}_n^N$ be the set of possible assignment vectors for $n \leq N$, that is,

$$\mathcal{Z}_n^N = \{(z_1, \ldots, z_N) : z_i \in \{0, 1\} \text{ for } i = 1, \ldots, n \text{ and } z_i = * \text{ for } i = n+1, \ldots, N\}.$$

For example, $\mathcal{Z}_2^4 = \{(0, 0, *, *), (0, 1, *, *), (1, 0, *, *), (1, 1, *, *)\}$.

The primary outcome of interest is whether or not a unit contracts HIV. Let $Y_i^t(\mathbf{z})$ be the HIV infection status of unit $i = 1, \ldots, N$ at time $t \in [0, T]$ under treatment assignment vector $\mathbf{z} \in \mathcal{Z}_n^N$. The time $t = 0$ corresponds to the beginning of the experiment (for all units) and T is a fixed point in time denoting the end of the experiment.

In addition to the primary outcome, sexual contacts occur after treatment is assigned and are a well defined outcome. Let $C_{i,i'}^t(\mathbf{z})$ be the total number of unprotected sexual contacts between units $i = 1, \ldots, N$ and $i' = 1, \ldots, N$ up until time $t \in [0, T]$ under treatment assignment vector $\mathbf{z} \in \mathcal{Z}_n^N$. We treat contacts symmetrically—$C_{i,i'}^t(\mathbf{z}) = C_{i',i}^t(\mathbf{z})$—but the notation is general enough to allow contacts to be defined asymmetrically (e.g., defining contact "directionally" as unit i contacts unit i'). By convention, define $C_{i,i'}^0(\mathbf{z}) = 0$ and $C_{i,i}^t(\mathbf{z}) = 0$ for all possible i, i', t, and $\mathbf{z}$. The contact process $C_{i,i'}^t(\mathbf{z})$ is a step function with respect to t that increments by one at the time of each contact. We define $C_{i,i'}^t(\mathbf{z})$ to be right continuous with left limits and assume no two contacts occur at exactly the same time.

With definitions of infection and a contact process, we can define exposure between susceptible unit i and infectious unit $i' \neq i$ with respect to the infection status of two units and the contact process. At time t and under treatment assignment $\mathbf{z}$, unit i is exposed by unit i' if unit i is susceptible immediately before time t, unit i' is infected

at the time t (and thus infectious), and units i and i' have unprotected sex at time t; that is $\lim_{s\uparrow t} Y_i^s(\mathbf{z}) = 0$, $Y_{i'}^t(\mathbf{z}) = 1$, and $C_{i,i'}^t(\mathbf{z}) = \lim_{s\uparrow t} C_{ij}^s(\mathbf{z}) + 1$. If unit i' infects unit i with exposure at time t, then $Y_i^t(\mathbf{z}) = 1$.

With the definition of a single exposure, we can count the number of times unit i is exposed by unit i'. Let $D_{i,i'}^t(\mathbf{z})$ be the number of times unit i' exposes unit i from time 0 to t under treatment assignment $\mathbf{z}$. This definition is not symmetric in that unit i' exposing unit i is not the same as unit i exposing unit i'. We can represent this quantity as

$$D_{i,i'}^t(\mathbf{z}) = C_{i,i'}^u(\mathbf{z}) - \lim_{s\uparrow l} C_{i,i'}^s(\mathbf{z})$$

where

$$u = \sup\{s \in [0,t] : Y_i^s(\mathbf{z}) = 0 \text{ and } Y_{i'}^s(\mathbf{z}) = 1\}$$

$$l = \inf\{s \in [0,t] : Y_i^s(\mathbf{z}) = 0 \text{ and } Y_{i'}^s(\mathbf{z}) = 1\}$$

and the notation for u and l suppress the obvious dependence on i, i', t, and $\mathbf{z}$ for simplicity. Lastly, we let $D_i^t(\mathbf{z})$ be the total number of exposures for unit i up to time t under treatment assignment $\mathbf{z}$, that is

$$D_i^t(\mathbf{z}) = \sum_{i'=1}^{N} D_{i,i'}^t(\mathbf{z}).$$

With the potential outcomes notation, causal effects are comparisons of the potential outcomes under the two different treatment assignment vectors. We define causal estimands more precisely in the next two sections under two different scenarios. In

Section 3.4, we consider a monogamous couples design, which avoids the complications of interference by design. We consider interference in Section 3.5 which discusses a closed population design in which all units can interact with each other.

3.4 Couples design

A couples design enrolls uninfected units who are in a monogamous relationship with one other partner. The HIV status of the partner is not necessarily known. If the HIV status of the partner is known to be infected, this study design is called a "discordant couples" design. If interference between units only occurs through sexual contacts, the couples design avoids the complications of interference. In the absence of interference, the potential outcomes notation can be simplified and we can focus on the issue of exposure to HIV.

3.4.1 Implications for potential outcomes

We consider an experimental population of n units who are not infected and are in a monogamous relationship over the course of the study. The non-experimental population consists of the n partners in each relationship. For notational convenience, we index the partner of experimental unit i as $i + n$. The simplifying assumption for the couples design is that each unit only has sexual contact with his partner, which obviates the problem of interference between units. The couples design implies the plausibility of the stable unit treatment value assumption (SUTVA Rubin, 1980), which states the following:

Assumption 3 (SUTVA) *There is no interference between units and there are no hidden versions of treatments. That is, for any* $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}_n^N$*, if* $z_i = z_i'$ *for* $i = 1, \ldots, N$*, then* $Y_i^t(\mathbf{z}) = Y_i^t(\mathbf{z}')$ *for all* $t \in [0, T]$*.*

SUTVA allows us to simplify the potential outcomes notation so that a unit's potential outcome is only a function of the treatment assignment indicator of that particular unit instead of the entire vector of treatment assignments. For this section, we use the notation $Y_i^t(1)$ and $Y_i^t(0)$ to be the two potential outcomes of HIV status for experimental unit $i = 1, \ldots, n$ at time t under the active treatment and control treatment respectively. Analogous notation applies to the number of contacts, $C_{i,i+n}^t(1)$ and $C_{i,i+n}^t(0)$ are the number of contacts between unit i and partner $i + n$ at time t when unit i is assigned to the active and control treatment respectively. The monogamous couples design implies $C_{i,i'}^t(z) = 0$ for $i' \neq i + n$, $t \in [0, T]$, and $z \in \{0, 1\}$. The simplified notation also applies to exposures, that is $D_{i,i+n}^t(1)$ and $D_{i,i+n}^t(0)$ are the number of times partner $i + n$ exposes unit i up until time t under the active and control treatments respectively.

Implicit in our assumptions is that the non-experimental population is unaffected by treatment assignments. For non-experimental unit $i = n + 1, \ldots, 2n$, the implication is that $Y_i^t(*)$ is constant with respect to t because unit i is not exposed to HIV through any contacts. Therefore, in a monogamous couples design, the infectiousness of the partner remains constant.

3.4.2 Identical exposure

Equipped with the potential outcomes notation, we return to the identical exposure condition of Greenwood and Yule (1915). We consider identical exposure based on assumptions of the contact process and the infectiousness of a partner because exposure is defined with respect to these concepts.

First, we define the identical contacts assumption, which states the following:

Assumption 4 (Identical contacts) *Treatment has no effect on contacts. In the general potential outcomes notation, for all* $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}_n^N$, $C_{ij}^t(\mathbf{z}) = C_{ij}^t(\mathbf{z}')$ *for all* $t \in [0, T]$.

For the monogamous couples design, this assumption states $C_{i,i+n}^t(1) = C_{i,i+n}^t(0)$ for all $i = 1, \ldots, n$ and $t \in [0, T]$. The interpretation of this assumption is that every unit in the population has the same sexual behavior regardless of the treatment assigned. Although defining identical infectiousness in the general case is difficult, identical infectiousness follows immediately from the couples design because the infectiousness of each experimental unit's partner is constant throughout the trial.

Even with an assumption of identical contacts and identical infectiousness in a couples design, the notion of identical exposure is still difficult to define. To see why, suppose the treatment is effective at reducing the susceptibility of the disease for all units. Then every unit under the treatment stays uninfected and susceptible longer. Denote the length of time a unit is susceptible as $L_i^t(\mathbf{z})$ under treatment assignment

vector $\mathbf{z}$, that is

$$L_i(\mathbf{z}) = \begin{cases} \inf\{t \in (0,T) : Y_i^t(\mathbf{z}) = 1\} & \text{if } Y_i^T(\mathbf{z}) = 1 \\ T & \text{otherwise.} \end{cases}$$

Using the simplified notation under the couples design, an effective treatment for experimental unit i implies $L_i(1) \geq L_i(0)$, which implies $D^T_{i,i+n}(1) \geq D^T_{i,i+n}(0)$ because $C^t_{i,i+n}(0) = C^t_{i,i+n}(1)$ is an increasing function in t. Therefore, an effective treatment on susceptibility necessarily implies an increase in the number of exposures. The identical exposure condition can be violated with an effective treatment. The simple intuition is that if the treatment effectively prevents infection for a given unit, then that unit will receive more exposures under the active treatment because the unit will become infected earlier under the control.

Defining estimands

Causal effects are comparisons of the potential outcomes under the two treatments for a common set of units. We consider estimands for the monogamous couples design and contrast ideas that use information on exposure and those that do not. We use contrasts that are differences, but other possible estimands could use different contrasts like ratios.

We define the *average causal effect on cumulative incidence* as

$$\tau_{\text{ci}} = \frac{1}{n}\sum_{i=1}^{n}[Y_i^T(1) - Y_i^T(0)] = \overline{Y}(1) - \overline{Y}(0),$$

where

$$\overline{Y}(z) = \frac{1}{n}\sum_{i=1}^{n} Y_i^T(z)$$

for $z \in \{0, 1\}$. The effect on cumulative incidence is not a function of exposure[1]. This estimand may not be scientifically interesting because we know that HIV is only transmitted to units that are exposed to the disease. Thus, an estimand that *is* a function of exposure is more relevant.

An estimand that is a function of exposure might compare the cumulative incidence for the exposed. We define an estimand that is specific to the couples design because the exposed subpopulation remains constant for this design. We define the *average causal effect on cumulative incidence for the exposed* as

$$\tau_{\text{cife}} = \frac{\sum_{i=1:Y_{i+n}^0(*)=1}^{n} [Y_i^T(1) - Y_i^T(0)]}{\sum_{i=1:Y_{i+n}^0(*)=1}^{n} 1}.$$

This estimand is more relevant to the problem of assessing the effect on susceptibility of the disease because is considers the subpopulation of experimental units that are exposed to the disease and actually have a chance of becoming infected.

An estimand that is analogous to the estimator in Halloran et al. (page 359, 1994) is the *difference of infections per exposure*:

$$\tau_{\text{ipe}} = \frac{\sum_{i=1}^{n} Y_i^T(1)}{\sum_{i=1}^{n} D_{i,i+n}^T(1)} - \frac{\sum_{i=1}^{n} Y_i^T(0)}{\sum_{i=1}^{n} D_{i,i+n}^T(0)}.$$

This estimand does not necessarily have a causal interpretation because potential

[1]We describe this estimand as (not a) function of exposure instead of "(un)conditional on exposure" as described in Rhodes et al. (1996) in order to differentiate between the potential outcomes and the model that is assumed or data that are collected.

outcomes are not being compared for each unit. We can interpret this estimand as treating each exposure as a unit and considering whether or not an infection occurs at each exposure. Recall that the number of exposures can differ under treatment and control (even under identical contacts and identical infectiousness), so this interpretation is not a causal effect because the set of exposures under treatment can be different from the set of exposures under control. This detail can be obviated by assuming all exposures are exchangeable, and then the exposures under treatment are assumed to be comparable to exposures under the control. Assuming symmetry on the exposures is a strong assumption. For instance, the right-most term will be the same under the following two scenarios under the control treatment: (1) only one unit is exposed and becomes infected after ten exposures and (2) five units are exposed twice and one unit is infected on the second exposure. Even if exposures are not truly exchangeable, the estimand that compares infections per exposure is a useful description of the effect of a prophylactic on HIV transmissions that reflects the general interpretation of how disease is transmitted. For example, τ_{ipe} can be interpreted as the difference in the probability of contracting the disease when exposed—the secondary attack rate—under the active and control treatment.

3.5 Closed population design

We now consider an experimental design in which interference can occur between units. The experimental population is composed of all units that are not infected at the beginning of the trial and the non-experimental population is composed of all units that are infected at the beginning of the trial. After treatment is assigned,

all units can contact each other. We call this design a "closed population" design because all possible contacts occur within the finite population for the duration of the trial, so interference only occurs between units in this finite population.

In this section, we first describe the causal estimands in the presence of interference as defined in HH. Next, we introduce a simple threshold model for the number of exposures until infection occurs. Combined with a sexual network of contacts, this model produces fixed potential outcomes for every treatment assignment vector. It additionally allows for simple description of the treatment effect on a unit's threshold. We close the section with an examination of the repercussions of this model on the estimands of HH in a simulated network of sexual contacts.

3.5.1 Causal estimands with interference

Interference between units implies that different treatment assignment vectors (e.g., $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}_n^N$) can have different outcomes for an experimental unit even if that unit is assigned the same treatment (e.g., $z_i = z_i'$, but $Y_i(\mathbf{z}) \neq Y_i(\mathbf{z}')$ for some $i \leq n$). HH apply two strategies when summarizing treatment effects in the presence of interference. First, for a given unit, HH average the potential outcomes over a specified intervention program, which dictates how treatments are assigned to the experimental population. Secondly, HH compare these average potential outcomes within and across intervention programs. The treatment for a each unit becomes a two factor treatment: the first factor determines the intervention program for the population and the second factor determines the active or control treatment for each unit.

To make this idea more concrete, we follow HH and consider two intervention programs that randomly assign a specified proportion of the experimental population to the active treatment. Denote the two programs as ψ and ϕ, which denote the proportion of experimental units assigned to the active treatment under intervention program for $\psi, \phi \in \{0, 1/n, 2/n, \ldots, n/n\}$ and $\psi \neq \phi$.

An experimental design that exploits the relationship between these two treatment factors is the split plot or pseudo cluster randomization. In these designs, first the intervention program is randomly assigned to a population and then given the first randomization, the second stage randomly assigns the unit level treatments following the randomly assigned intervention program. We do not discuss further inference for these designs because our focus is on quantifying interference through definitions of causal estimands (see HH for details on inference).

Average potential outcomes

With the additional treatment factor that specifies the intervention program, HH summarize the potential outcomes for unit i under assignment to treatment z and specified intervention program ψ by averaging all potential outcomes with respect to the intervention program given unit i is assigned to treatment z. That is, define the *individual average potential outcome* for experimental unit $i = 1, \ldots, n$, under treatment $z \in \{0, 1\}$ for intervention program ψ as

$$\overline{Y}_i(z, \psi) \equiv \sum_{\mathbf{z} \in \mathcal{Z}_n^N} Y_i^T(\mathbf{z}) \Pr_\psi(\mathbf{Z} = \mathbf{z} | Z_i = z).$$

The unit level average potential outcome can be summarized by averaging over

all experimental units in the experimental population. That is, define the *population average potential outcome* under treatment z for intervention program ψ as

$$\overline{Y}(z,\psi) \equiv \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i(z,\psi).$$

The individual average potential outcomes are defined for a intervention program and treatment assignment, but we could also consider an average potential outcome that does not constrain the treatment assignment of the unit. Define the *marginal individual average potential outcome* for experimental unit $i = 1,\ldots,n$ under intervention program ψ as

$$\overline{Y}_i(\psi) \equiv \sum_{\mathbf{z}\in\mathcal{Z}_n^N} Y_i^T(\mathbf{z})\Pr_\psi(\mathbf{Z}=\mathbf{z}).$$

As before, the unit level averages can be summarized by averaging over all experimental units. Define the *marginal population average potential outcome* under intervention program ψ as

$$\overline{Y}(\psi) = \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i(\psi).$$

Direct, indirect, total, and overall causal effects

From the population average potential outcomes, HH define direct, indirect, total, and overall causal effects. The following definitions of direct and indirect effects are causal estimands in the presence of interference and should not be confused with similar terminology of direct and indirect effects associated with causal graphs (e.g., Pearl, 1995).

The *population average direct causal effect* for intervention program ψ is the dif-

ference of the population average potential outcomes under the active and control treatments for intervention program ψ, that is

$$\overline{DCE}_Y(\psi) \equiv \overline{Y}(1,\psi) - \overline{Y}(0,\psi).$$

Intuitively, the direct effect compares taking the active treatment to the control treatment under the same intervention program. Thus, it is assessing the "direct" effect of the experimental treatment compared to the control treatment.

The *population average indirect causal effect* for intervention program ψ compared to intervention program ϕ is the difference of the population average potential outcomes under the control treatment for the two intervention programs ψ and ϕ, that is

$$\overline{ICE}_Y(\psi,\phi) \equiv \overline{Y}(0,\psi) - \overline{Y}(0,\phi).$$

Intuitively, the indirect effect compares taking the control treatment under two different intervention programs. Thus, any effect on the outcomes is due to an "indirect" effect from the intervention program and not from taking the active treatment.

The *population average total causal effect* for intervention program ψ compared to intervention program ϕ is the difference of the population average potential outcomes under receipt of the active treatment in intervention program ψ and the average of the potential outcomes under receipt of the control treatment in intervention program ϕ, that is

$$\overline{TCE}_Y(\psi,\phi) = \overline{Y}(1,\psi) - \overline{Y}(0,\phi).$$

Intuitively, the total effect compares taking the active treatment under intervention

program ψ to taking the control treatment under intervention program ϕ. Thus, any effect on the outcome is due to the combined or "total" effect of taking the active treatment and the intervention program compared to the control treatment and some control intervention program. The total effects is the sum of the indirect and direct effects: $\overline{TCE}_Y(\psi, \phi) = \overline{DCE}_Y(\psi) + \overline{ICE}_Y(\psi, \phi)$.

Lastly, the *population average overall effect* for intervention program ψ compared to intervention program ϕ is the difference of the marginal population average potential outcomes for the two intervention programs, that is

$$\overline{OCE}_Y(\psi, \phi) = \overline{Y}(\psi) - \overline{Y}(\phi).$$

Intuitively, the overall effect compares the number of infections in the population under the two different intervention programs and ignores who actually receives the treatment. Thus, the effect is an "overall" comparison of the two different treatment regimes.

Example 3.5.1 *We demonstrate the computation of these effects for a finite population of $N = 3$ units and experimental population of $n = 2$ units. We let $\psi = 1/2$ and $\phi = 0$. The outcomes for the three assignment vectors are provided in Table 3.1.*

We have

$$\overline{DCE}_Y(\psi) = 0 - 1/2 = -1/2,$$

$$\overline{ICE}_Y(\psi, \phi) = 1/2 - 1 = -1/2,$$

$$\overline{TCE}_Y(\psi, \phi) = 0 - 1 = -1,$$

$$\overline{OCE}_Y(\psi, \phi) = 1/4 - 1 = -3/4.$$

Table 3.1: Outcomes for example population under three treatment assignment vectors. The first treatment assignment vector corresponds to $\phi = 0$; the next two treatment assignment vectors correspond to $\psi = 1/2$.

		t	
$\mathbf{z}$		0	T
$(0, 0, *)$	$Y_1^t(\mathbf{z})$	0	1
	$Y_2^t(\mathbf{z})$	0	1
	$Y_3^t(\mathbf{z})$	1	1
$(0, 1, *)$	$Y_1^t(\mathbf{z})$	0	0
	$Y_2^t(\mathbf{z})$	0	0
	$Y_3^t(\mathbf{z})$	1	1
$(1, 0, *)$	$Y_1^t(\mathbf{z})$	0	0
	$Y_2^t(\mathbf{z})$	0	1
	$Y_3^t(\mathbf{z})$	1	1

We defined the direct, indirect, total, and overall causal effects on infection. We later consider these effects on the amount of exposure for each unit. Using analogous notation (i.e., substitute $D_i^T(\mathbf{z})$ for $Y_i^T(\mathbf{z})$), let $\overline{DCE}_D(\psi)$, $\overline{ICE}_D(\psi, \phi)$, $\overline{TCE}_D(\psi, \phi)$, and $\overline{OCE}_D(\psi, \phi)$ be the population average direct, indirect, total, and overall causal effects on total exposure.

3.5.2 Threshold model for HIV transmission

The estimands from HH rely on fixed potential outcomes for each treatment assignment vector. We propose a threshold model for HIV transmission that produces fixed outcomes of HIV infection for every treatment assignment vector in combination with the network of sexual contacts. We make the simplifying assumption of identical contacts, and suppress the dependence of the network on the treatment assignment vector.

The threshold model assumes that for a given unit there is a fixed number of exposures to HIV such that the unit becomes infected once he is exposed as many times as his threshold. We assume all exposures count the same towards the total number of exposures. That is, exposures at different times and from different units all increment the number of exposures. Additionally, we make stable unit treatment value assumption for the threshold values, so we can consider the threshold under assignment to the active treatment and control treatment. As discussed in Halloran and Struchiner (1995), considering the potential outcomes as functions of exposures "seems to win back the stability assumption." We let $M_i(z)$ be the threshold value for experimental unit i under assignment to treatment $z \in \{0, 1\}$.

Under SUTVA for the threshold model, interference on infections is due to interference on the number of exposures. That is, suppose we have two treatment assignment vectors $\mathbf{z}$ and $\mathbf{z}'$ where $z_i = z'_i$. Unit i becomes infected under the treatment assignment vector if the number of exposures reaches the threshold, but the number of exposures may be different under the two treatment assignment vectors. For $\mathbf{z}$ and $\mathbf{z}'$, an example of interference occurring is $D_i^T(\mathbf{z}) < D_i^T(\mathbf{z}') = M_i(z_i)$, in

which case $Y_i^T(\mathbf{z}) = 0 \neq Y_i^T(\mathbf{z}') = 1$.

The threshold model provides a simple way to describe the effect of treatment on a unit. For example, if $M_i(1) > M_i(0)$, then the treatment has an effect of preventing infection for unit i. We can also consider homogeneous effects across units, for example the multiplicative effect $M_i(1) = \lceil \theta M_i(0) \rceil$ where $\theta \geq 0$ and $\lceil x \rceil$ is the smallest integer greater than x; an effective treatment would correspond to $\theta > 1$. We call such effects *threshold treatment effects* to distinguish them from our previous estimands. We return to the simple example with three units to demonstrate the threshold model.

Example 3.5.2 *We return to the example of a finite population of $N = 3$ units and and experimental population of $n = 2$ units. The experiment occurs over time $T = 5$. The contact process (under the assumption of identical contacts) is provided in the first three rows of a Table 3.2. We assume the $M_1(0) = M_2(0) = 2$ and $M_1(1) = M_2(1) = 4$. The potential outcomes of HIV infection are also provided in Table 3.2 for three different treatment assignment vectors.*

Under treatment assignment vector $\mathbf{z} = (0, 0, *)$, unit 2 is exposed twice by unit 3 and contracts HIV on the second exposure because $M_2(0) = 2$. Unit 2 then exposes unit 1, who contracts HIV on the second exposure. Both units in the experimental population contract HIV under this treatment assignment.

Under treatment assignment vector $\mathbf{z} = (0, 1, *)$, unit 2 is exposed twice by unit 3 but does not contract HIV because he does not reach his threshold of $M_2(1) = 4$. Unit 1 is not exposed at all and therefore does not contract HIV.

Under treatment assignment vector $\mathbf{z} = (1, 0, *)$, unit 2 is exposed twice by unit 3

Table 3.2: Example of the threshold model for $M_1(0) = M_2(0) = 2$ and $M_1(1) = M_2(1) = 4$.

		t					
$\mathbf{z}$		0	1	2	3	4	5
	$C^t_{1,2}$	0	0	0	1	2	3
	$C^t_{2,3}$	0	1	2	2	2	2
	$C^t_{1,3}$	0	0	0	0	0	0
$(0,0,*)$	$Y^t_1(\mathbf{z})$	0	0	0	0	1	1
	$Y^t_2(\mathbf{z})$	0	0	1	1	1	1
	$Y^t_3(\mathbf{z})$	1	1	1	1	1	1
$(0,1,*)$	$Y^t_1(\mathbf{z})$	0	0	0	0	0	0
	$Y^t_2(\mathbf{z})$	0	0	0	0	0	0
	$Y^t_3(\mathbf{z})$	1	1	1	1	1	1
$(1,0,*)$	$Y^t_1(\mathbf{z})$	0	0	0	0	0	0
	$Y^t_2(\mathbf{z})$	0	0	1	1	1	1
	$Y^t_3(\mathbf{z})$	1	1	1	1	1	1

and does contract HIV. Unit 2 exposes unit 1 three times, but unit 1 does not contract HIV because he did not reach his threshold of $M_1(1) = 4$.

Although the threshold model provides a simple description of the treatment effects, the ramifications for the estimands of HH are not obvious. This simple example demonstrates how to incorporate the threshold model with the network of sexual contacts to produce fixed potential outcomes for each treatment assignment vector.

The threshold model is not probabilistic. $M_i(z)$ can be thought of as a latent variable describing a unit's innate susceptibility to HIV under assignment to treatment z. Assuming a probability model for $M_i(z)$ leads to standard models for disease transmission. For example, if $M_i(z) \sim \text{Geometric}(\pi_z)$, we can interpret exposures as independent Bernoulli trials with probability of success (i.e., disease transmission) π_z. The assumptions of this model are similar to those for the estimand τ_{ipe}, which is

the finite population analog of comparing the parameters π_1 to π_0. In our simulation study, we simulate $M_i(z)$ using the geometric distribution.

3.5.3 Assessing interference on a simulated closed population study

We provide a simulation of a closed population design in order to demonstrate the relationship between direct, indirect, total, and overall causal effects on both infections and exposures. We compare various values of intervention program ψ to a "control" intervention program $\phi = 0$, which corresponds to no one receiving the treatment. We also consider various values of the threshold treatment effect as described using the threshold model with a multiplicative effect. Our simulations show that the amount of interference is a function of the intervention program and the threshold treatment effect. The important take away from examining all four estimands is that there is potential for dramatic variability due to interference. We first describe the simulation and then summarize the results.

Simulated closed population

The experimental population is composed of $n = 5000$ uninfected units and the non-experimental population is composed of $N - n = 555$ infected units. That is, 10% of the population is HIV infected at the beginning of the trial. Initial infection status of each unit is simulated once, independent of all the other quantities. The length of the study is three years.

We simulate the network of contacts once under the assumption of identical con-

tacts. Details of the how we simulate the sexual contacts are provided in Appendix C. An accurate simulation of a sexual network is a difficult task, but we attempted to replicate some descriptive statistics from observed sexual networks in the literature. Summaries of the network are provided in Table 3.3.

Table 3.3: Summaries statistics of the network of sexual contacts. The degree distribution summarizes the number of sexual partners per unit. The network summaries describe the "connectedness" of the network (compare to similar summaries in Helleringer and Kohler, 2007). The last table summarizes the number of sexual contacts per unit over the course of the study.

Degree distribution	
Degree	Value
1	0.62
2	0.24
3	0.06
4	0.03
> 4	0.05

Network summaries	
Description	Value
% of clusters of size < 5	82%
% of population in clusters of size < 5	34%
% of population in largest cluster	47%

Sexual contacts per unit	
Description	Value
Mean	169
Standard deviation	149
Median	127
25th percentile	36
75th percentile	300

Given the network of sexual contacts, we use the threshold model to generate fixed potential outcomes for each treatment assignment. We simulate the threshold value under the control treatment using the geometric distribution with probability of transmission $\pi = 0.02$. The threshold value under the alternative varies according to the multiplicative threshold treatment effect of $\theta \in \{2, 4, 8, 16\}$.

With the exception of the "control" intervention program with no units assigned to the active treatment, calculating the population average potential outcomes requires enumerating several treatment assignment vectors. For example, if $\psi = 1/2$, then there are $\binom{5000}{2500}$ treatment assignment vectors to consider. We do not enumerate all possible vectors, but instead perform a Monte Carlo simulation. We randomly sample 1000 treatment assignment vectors from each intervention program in order to estimate the finite population estimands.

Direct, indirect, total, and overall effects in simulated population

Given our closed population, we quantify the interference within a randomized experiment using the population average direct, indirect, total, and overall causal effects on infection, which are plotted in Figure 3.1 under different treatment intervention programs and different threshold treatment effects. As a point of reference, 12.3% of the experimental units are infected under the control intervention program. Thus, average effects (comparing intervention program to the control program) of -0.1 would be considered a huge success (i.e., 81% decrease), but effects of -0.01 could still be considered a modest success (i.e., 8% decrease).

Our first general observation is that there is large variability in these estimands across different treatment regimes and treatment effects. For a given threshold treatment effect, the trends of the average direct, indirect, and total causal effects as functions of the proportion treated is due solely to interference between units.

One intuitive observation is that the magnitude of the indirect average causal effect increases as the proportion of units that are treated increases. When only 5% of the

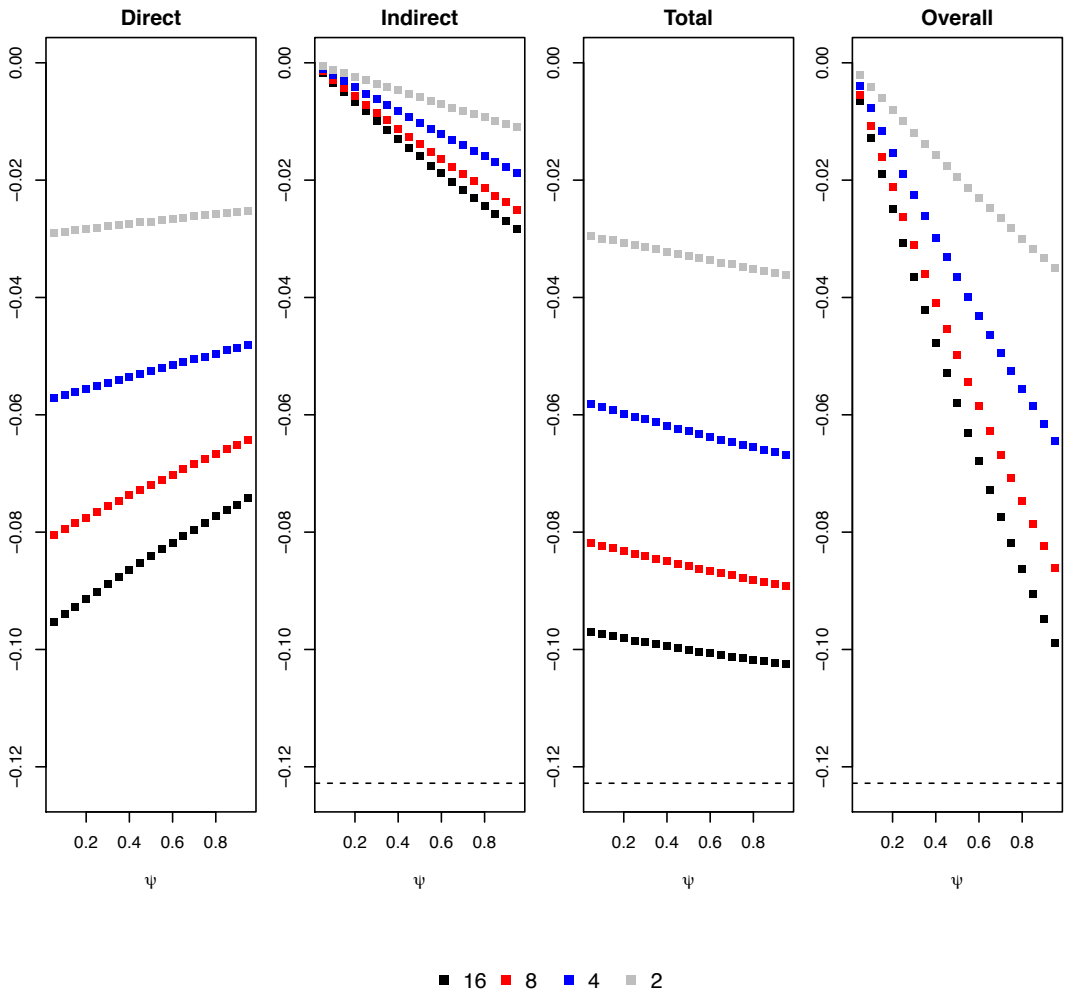

Figure 3.1: Direct, indirect, total, and overall average causal effects on proportion of experimental units that contract HIV. The y-axis corresponds to the magnitude and direction of the effect; the x-axis corresponds to the proportion of the experimental population that is treated. The different colors denote different multiplicative treatment effects for the threshold model. The dotted lines for indirect, total, and overall effects denote the maximum possible treatment effect.

units are assigned to treatment, there is essentially no indirect average causal effect, whereas when $\theta = 2$ the indirect effect is -0.005 and -0.010 when 50% and 90% of the population are randomly treated respectively. A large indirect effect corresponds to the concept of interference. Thus, interference increases as the proportion of the population that is treated increases and the effectiveness of the treatment increases.

Perhaps less intuitive is that the direct effect shrinks towards zero as the proportion of units that are treated increases. The explanation for this result is that the indirect effects are protecting units that would have been protected "directly" if assigned the active treatment. This explanation becomes clear when examining the total average causal effect, which is the sum of the direct and indirect effects. The magnitude of the total average causal effect shows a modest increase.

As pointed out in HH (Theorem 1), the comparison of observed infection rates in the active treatment and control treatment groups is an unbiased estimate of the direct effect. If no interference is assumed, then this estimator might be incorrectly interpreted as estimating the difference in cumulative incidence rates under treatment and control without acknowledging the potential benefits of interference (i.e., the indirect effect). Estimation of indirect, total, and overall effects requires either more populations that could be assigned different treatment regimes, or detailed knowledge of the network of contacts.

The direct, indirect, and total effects are descriptions of unit level effects. In contrast, the overall causal effect might be more relevant from the perspective of public health, because the overall effect summarizes effect of the intervention program, which could be implemented in other populations.

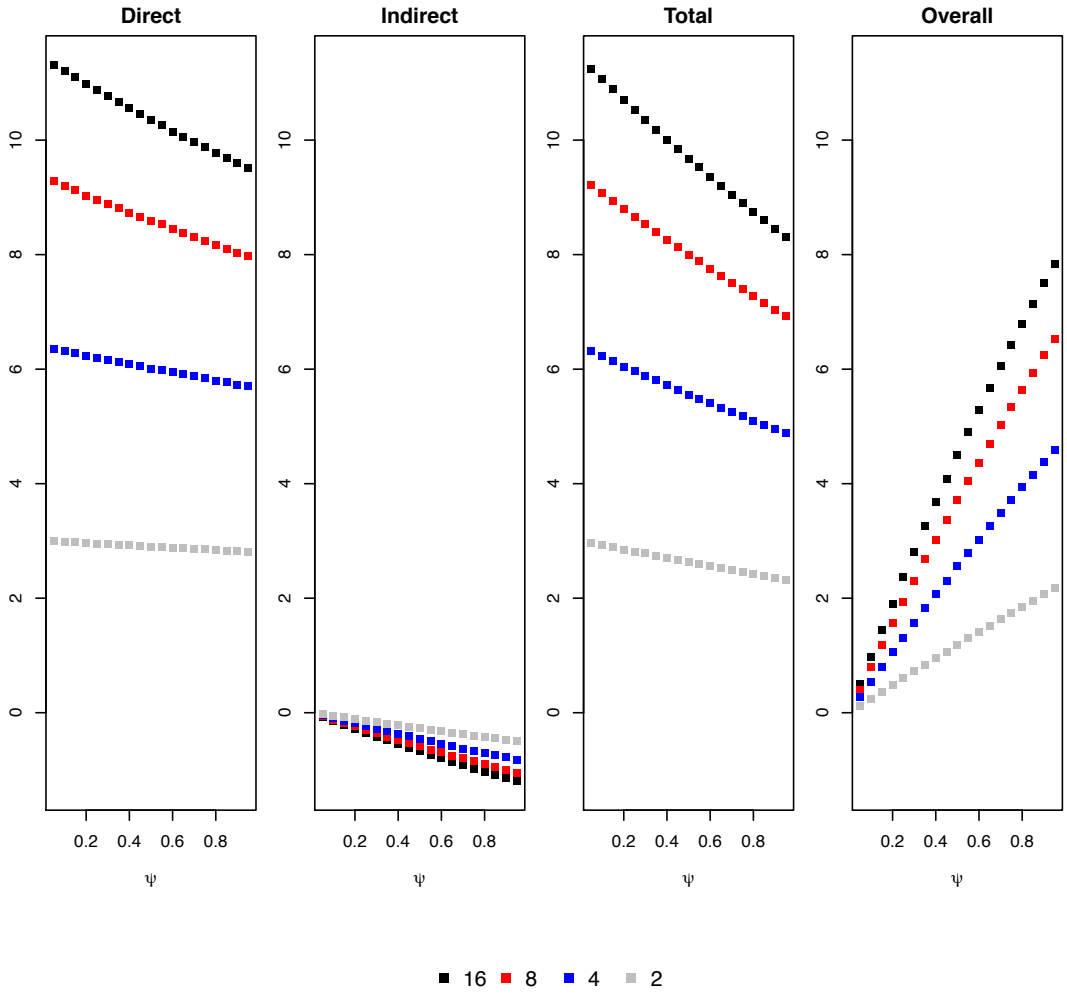

Figure 3.2: Direct, indirect, total, and overall average causal effects on the number of exposures. The y-axis corresponds to the magnitude and direction of the effect; negative effects correspond to a reduction in exposures. The x-axis corresponds to the proportion of the experimental population that is treated. The different colors denote different threshold treatment effects.

We also consider the direct, indirect, total, and overall average causal effects on the average number of exposures per unit. These effects are plotted in Figure 3.2. This plot demonstrates two points discussed previously. First, the positive direct average causal effect follow from the argument that an effective HIV prophylactic necessarily increases the number of exposures. Secondly, the negative indirect effect follows from the argument that assignment to the active treatment may prevent some units from contracting HIV and exposing other units to HIV. The modest size of the indirect effect is partly due to averaging over all units, a large number of whom are never exposed.

3.6 Conclusion

Assessing the efficacy of treatments for preventing infectious disease is a quintessential example of interference occurring between units. We discuss the role of exposure to disease as the reason for interference between units. In the case of a prophylactic treatment for HIV, interference can be described as occurring within a network of sexual contacts. Interference occurs because the treatment assignment for some units affects the degree to which they expose other units to HIV.

We demonstrate these concepts in a simulated population using a threshold model for the number of exposures until infection. Our simulation study shows the relationship between the direct, indirect, and total causal effects that are defined in HH. The simulation study shows intuitive results about the amount of interference depending on the effectiveness of the treatment as well as the proportion of units in the population that are treated. Perhaps less intuitive are the results about direct effects

attenuating as the proportion of treated subjects increases due to the indirect effect of the treatment. Our results contribute to the understanding of causal inference in the presence of interference for studies of prophylactic treatments for infectious diseases by examining a the specific mechanism by which HIV is transmitted.

Appendix A

Supplement to Chapter 1

A.1 Estimand for analysis that discards rescued subjects

We consider the estimator that discards rescued subjects $\hat{\phi}_{\text{dr}}$ as defined in Equation (1.2). We argue that for a fixed population of size N and randomly assigning N_{t} units

to treatment and the remaining $N_{\mathrm{p}} = N - N_{\mathrm{t}}$ units to placebo, $\mathrm{E}[\hat{\phi}_{\mathrm{dr}}] \approx \phi_{\mathrm{dr}}$ where

$$\begin{aligned}
\phi_{\mathrm{dr}} &= \frac{\mathrm{E}\left(\sum_{i\in\mathcal{O}(1,0)} Y_i^{\mathrm{obs}}\right)}{\mathrm{E}\left(\sum_{i\in\mathcal{O}(1,0)} 1\right)} - \frac{\mathrm{E}\left(\sum_{i\in\mathcal{O}(0,0)} Y_i^{\mathrm{obs}}\right)}{\mathrm{E}\left(\sum_{i\in\mathcal{O}(0,0)} 1\right)} \\
&= \frac{\sum_{i:S_i=\mathrm{NN}} \mathrm{E}[Z_i]Y_i(1) + \sum_{i:S_i=\mathrm{NR}} \mathrm{E}[Z_i]Y_i(1)}{\sum_{i:S_i=\mathrm{NN}} \mathrm{E}[Z_i] + \sum_{i:S_i=\mathrm{NR}} \mathrm{E}[Z_i]} \\
&\quad - \frac{\sum_{i:S_i=\mathrm{NN}} (1-\mathrm{E}[Z_i])Y_i(0) + \sum_{i:S_i=\mathrm{RN}} (1-\mathrm{E}[Z_i])Y_i(0)}{\sum_{i:S_i=\mathrm{NN}} (1-\mathrm{E}[Z_i]) + \sum_{i:S_i=\mathrm{RN}} (1-\mathrm{E}[Z_i])} \\
&= \frac{\pi_{\mathrm{NN}}\overline{Y}_{\mathrm{NN}}(1) + \pi_{\mathrm{NR}}\overline{Y}_{\mathrm{NR}}(1)}{\pi_{\mathrm{NN}} + \pi_{\mathrm{NR}}} - \frac{\pi_{\mathrm{NN}}\overline{Y}_{\mathrm{NN}}(0) + \pi_{\mathrm{RN}}\overline{Y}_{\mathrm{RN}}(0)}{\pi_{\mathrm{NN}} + \pi_{\mathrm{RN}}},
\end{aligned}$$

where the last line follows from noting $\mathrm{E}[Z_i] = N_{\mathrm{t}}/N$ for all $i = 1, \ldots, N$.

To see the approximation, consider the Taylor series expansion of $\mathrm{E}\left[\frac{x}{y}\right]$ at μ_x and μ_y, the mean of positive random variables x and y respectively:

$$\mathrm{E}\left[\frac{x}{y}\right] \approx \frac{\mu_x}{\mu_y} + \frac{\mu_x}{\mu_y^3}\mathrm{Var}(y) - \frac{1}{2\mu_y^2}\mathrm{Cov}(x,y).$$

We apply this approximation twice for the two ratios of $\hat{\phi}_{\mathrm{dr}}$. For the first term of $\hat{\phi}_{\mathrm{dr}}$, let

$$x = \frac{1}{N_{\mathrm{t}}} \sum_{i\in\mathcal{O}(1,0)} Y_i^{\mathrm{obs}} \quad \text{and} \quad y = \frac{1}{N_{\mathrm{t}}} \sum_{i\in\mathcal{O}(1,0)} 1.$$

The first term of the approximation corresponds to the first term of ϕ_{dr}. Moreover, because x and y^1 are sample means, $\mathrm{Var}(y)$ and $\mathrm{Cov}(x,y)$ are on the order of $\frac{1}{N_{\mathrm{t}}}$.

An analogous approximation to the second term of Equation (1.2) shows that $\hat{\phi}_{\text{dr}}$ is a biased estimator of ϕ_{dr} where the bias is on the order of $\max(\frac{1}{N_{\text{t}}}, \frac{1}{N_{\text{p}}})$. Thus, $\hat{\phi}_{\text{dr}}$ estimates the estimand ϕ_{dr} with a bias that goes to zero as N_{t} and N_{p} increase to infinity.

A.2 Imputing missing data

We describe the imputation of principal strata membership and missing potential outcome of the clinical endpoint for each subject, given a posterior draw of the parameters and the observed data. First, given the parameter $\boldsymbol{\theta}$, all subjects' principal strata membership and missing potential outcomes are independent of each other, so we focus on imputing the missing data for subject $i = 1, \ldots, N$. We first draw the principal stratum membership S_i, and then, conditional on S_i, we draw the missing potential outcome $Y_i(1 - Z_i)$. We suppress the conditioning on the observed data and parameters for conciseness; instead we use subscript "post" to denote this conditional probability.

We consider the four possible observed combinations of assigned treatment (Z_i) and observed indicator of receipt of rescue medication (D_i^{obs}). If $i \in \mathcal{O}(0, 0)$, then

$$\Pr_{\text{post}}(S_i = s) = \begin{cases} 1 & \text{if } s = \text{NN} \\ 0 & \text{otherwise.} \end{cases}$$

[1]In complete generality, either denominator has a non-zero probability of being zero, and thus the mean of $\hat{\phi}_{\text{dr}}$ does not always exist. In the context of rescue medication, if the number of never rescued subjects is larger than both N_{t} and N_{p}, that is $N\pi_{\text{NN}} > \max(N_{\text{t}}, N_{\text{p}})$, then there is zero probability of either denominator being zero.

If $i \in \mathcal{O}(1,1)$, then

$$\Pr_{\text{post}}(S_i = s) = \begin{cases} 1 & \text{if } s = \text{RR} \\ 0 & \text{otherwise.} \end{cases}$$

If $i \in \mathcal{O}(0,1)$, then

$$\Pr_{\text{post}}(S_i = s) \propto \begin{cases} \pi_{\text{RR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{RR},i}) & \text{if } s = \text{RR} \\ \pi_{\text{NR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{NR},i}) & \text{if } s = \text{NR} \\ 0 & \text{otherwise.} \end{cases}$$

If $i \in \mathcal{O}(1,0)$, then

$$\Pr_{\text{post}}(S_i = s) \propto \begin{cases} \pi_{\text{NR},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{NR},i}) & \text{if } s = \text{NR} \\ \pi_{\text{NN},i} \cdot Poi(Y_i^{\text{obs}}; \lambda_{0,\text{NN},i}) & \text{if } s = \text{NN} \\ 0 & \text{otherwise.} \end{cases}$$

Given $S_i = s$ and $Z_i = z$ for $s \in \{\text{RR}, \text{NR}, \text{NN}\}$ and $z \in \{0,1\}$, we draw the missing potential outcome $Y_i(1-z)$ from a Poisson distribution with mean $\lambda_{1-z,s,i}$. By our conditional independence assumption on the joint distribution of the potential outcomes $Y_i(0)$ and $Y_i(1)$, the missing potential outcome does not depend on the observed potential outcome given the principal strata membership.

A.3 Checking software using prior and posterior simulations

We ensure that our implementation of our Bayesian sampler is correct using the simulation-based technique described in Cook et al. (2006). The procedure samples one draw of the parameter, say θ_0, from the (proper) prior distribution (for now, suppose θ is one dimensional). Conditional on θ_0, sample the data, say generically y, from the likelihood. Given the data, y, the software that is being checked generates several draws from the posterior distribution of the parameters, say $\theta_1, \ldots, \theta_R$ for R replicates (or perhaps iterations). The key idea is that θ_0 is a valid draw from the posterior distribution. Therefore, if the software is correctly sampling from the posterior distribution of the parameters, then $\theta_1, \ldots, \theta_R$ and θ_0 have the same distribution. To assess whether or not θ_0 and $\theta_1, \ldots, \theta_R$ come from the same posterior distribution, we repeat the process and note that the random variable

$$q = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\{\theta_0 \leq \theta_r\}$$

will have an approximate uniform distribution under repeated samples of the prior distribution and correct sampling of the posterior distribution (see Theorem 1 in Cook et al., 2006). For multidimensional θ, we calculate q for each component of θ.

We perform checks of the software implementing the sampler using the prior distribution describe in Section 5.3.2. We simulate 1000 replications of this procedure and generate a distribution of q for each of the 23 parameters. For each of the 23 parameters, we perform a Kolmogorov-Smirnov test of q (assuming a uniform null

distribution); p-values are presented in Table A.1. The Kolmogorov-Smirnov tests do no provide evidence against the null hypothesis that the code has a bug. Figure A.1 shows QQ-plots of the empirical quantiles of $\Phi^{-1}(q)$ against the expected quantiles of a standard normal distribution. If the software is performing correctly, $\Phi^{-1}(q)$ has a normal distribution. We use the probit transformation in order to accentuate the tails of the distribution. Except for some deviations in the tails of the distribution, the QQ-plots suggest no large deviations from normality and no evidence that the code has a bug. Therefore, we believe that the software is properly sampling from the posterior distribution of the parameters.

A.4 Model checking with posterior predictive distributions

We check to see if the model is consistent with observed data. The general idea is to compare aspects of the observed data to the posterior predictive distribution of a new set of data. Rubin et al. (1984) proposed the posterior predictive p-value, defined as

$$p_{\text{pp}} = \Pr(T(\text{data}^{\text{pp}}) \geq T(\text{data}) \mid \text{data}),$$

where data^{pp} represents the posterior predictive distribution of the data and T is a test statistic. Observing small (and perhaps large) values of p_{pp} correspond to a lack of fit for the model. We consider the posterior predictive distribution for the same set of units (i.e., $\mathbf{X}$ is held constant) and treatment assignment (i.e., $\mathbf{Z}$ is held constant). Gelman et al. (2003, Chapter 6) and Gelman (2003) discuss the use of

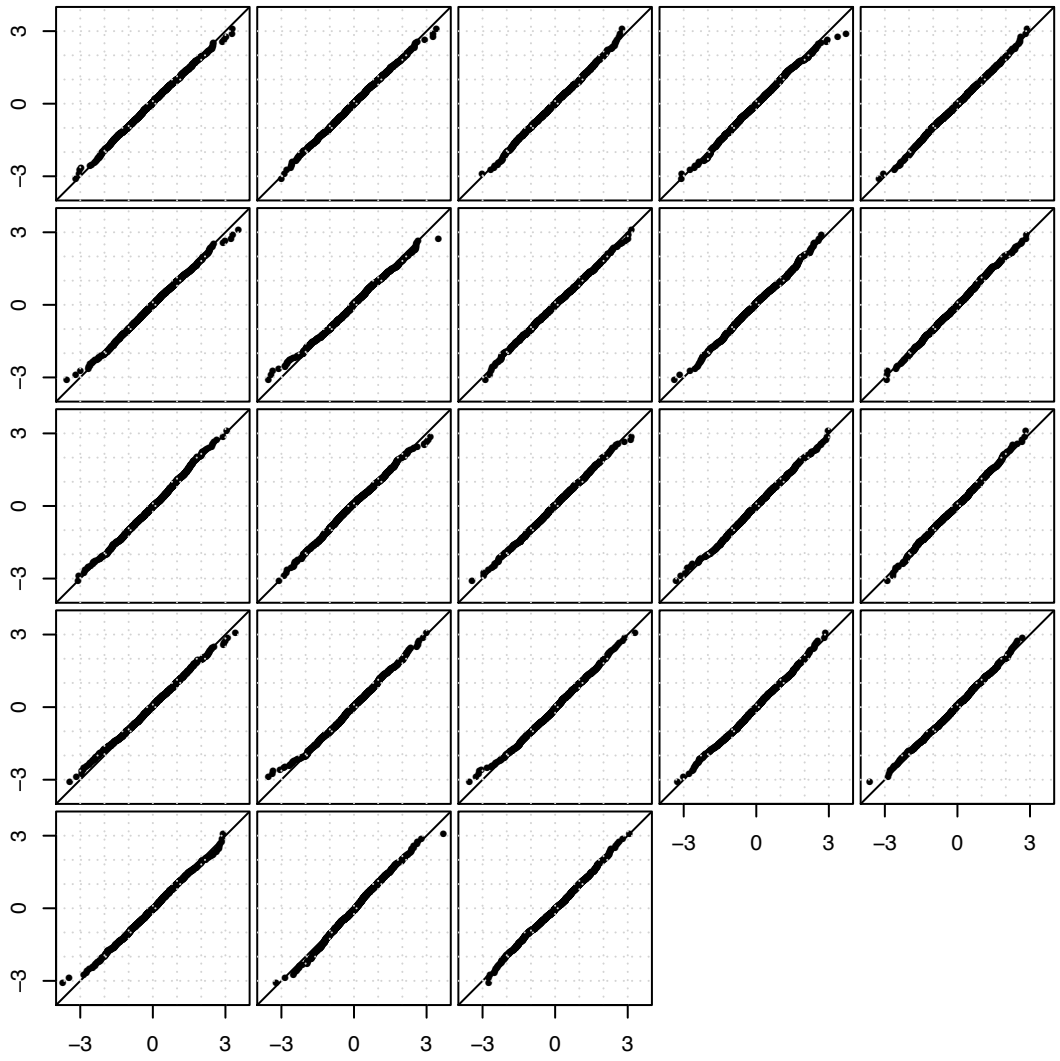

Figure A.1: QQ-plots of empirical quantiles of $\Phi^{-1}(q)$ and theoretical quantiles of the standard normal distribution (corresponding to the x-axis and y-axis respectively). Parameters for each plot correspond to the ordering in Table A.1 from right to left, top to bottom.

Table A.1: P-values from Kolmogorov-Smirnov test that q is distributed uniformly. Parentheses following subscripts of X indicate to which covariate the parameter corresponds.

Parameter	P-value
$\gamma_{0,\text{RR}}$	0.26
$\gamma_{1,\text{RR}}$	1.00
$\gamma_{0,\text{NR}}$	0.90
$\gamma_{1,\text{NR}}$	0.68
$\gamma_{0,\text{NN}}$	0.63
$\gamma_{1,\text{NN}}$	0.88
$\gamma_{\text{X(prior relapses)}}$	0.78
$\gamma_{\text{X(Gd lesions)}}$	0.34
$\gamma_{\text{X(T2 lesions)}}$	0.31
$\gamma_{\text{X(EDSS)}}$	0.83
$\gamma_{\text{X(Age)}}$	0.48
β_{NN}	0.03
β_{NR}	0.95
$\beta_{\text{NN,X(prior relapses)}}$	0.96
$\beta_{\text{NR,X(prior relapses)}}$	0.80
$\beta_{\text{NN,X(Gd lesions)}}$	0.49
$\beta_{\text{NR,X(Gd lesions)}}$	0.64
$\beta_{\text{NN,X(T2 lesions)}}$	0.05
$\beta_{\text{NR,X(T2 lesions)}}$	0.38
$\beta_{\text{NN,X(EDSS)}}$	0.29
$\beta_{\text{NR,X(EDSS)}}$	0.77
$\beta_{\text{NN,X(age)}}$	0.16
$\beta_{\text{NR,X(age)}}$	0.10

graphical summaries for the posterior predictive checks. In the Section A.4.1, we quantify p_{pp} for test statistics that target whether or not the data is over dispersed for the Poisson model. In the Section A.4.2 we consider graphical summaries that examine the relationship between the observed outcome and the observed indicator of rescue medication.

A.4.1 Numerical summaries

One concern with using a Poisson model is over dispersion. To examine whether or not the proposed model captures the correct amount of variability in the data, we consider two test statistics: the sample variance, $S^2 = \frac{1}{N-1}\sum \left(Y_i^{\text{obs}} - \overline{Y}^{\text{obs}}\right)^2$, and the maximum, $M = \max\{Y_i^{\text{obs}} : i = 1, \ldots, N\}$. Evidence of over dispersion would correspond to a posterior predictive distributions of these test statistics having most of the mass below the observed value the test statistics (i.e., small p_{pp}). The posterior predictive distributions of the test statistics along with the correspond posterior predictive p-value are provided in Figure A.2. These numerical summaries suggest that the Poisson model is capturing the variability in the data.

A.4.2 Graphical summaries

We also consider graphical summaries of the data to investigate the overall distribution of the data compared to the distribution of the predicted data. Our analysis focuses on the principal strata level effects, so we are particularly interested in examining the observed outcome distribution in relation to the observed indicator of rescue medication. Thus, we consider graphical summaries by the four combinations of treatment assignment and observed indicator of rescue medication. Figure A.3 compares the posterior predictive distributions against the distribution of the data for the observed outcomes via histograms. Lack of fit would be detected through dissimilar histograms for the posterior predictive distributions and the data distribution. Additionally, Figure A.4 shows QQ-plots of the posterior predictive distribution against the data distribution; points are jittered to handle ties. Lack of fit would manifest

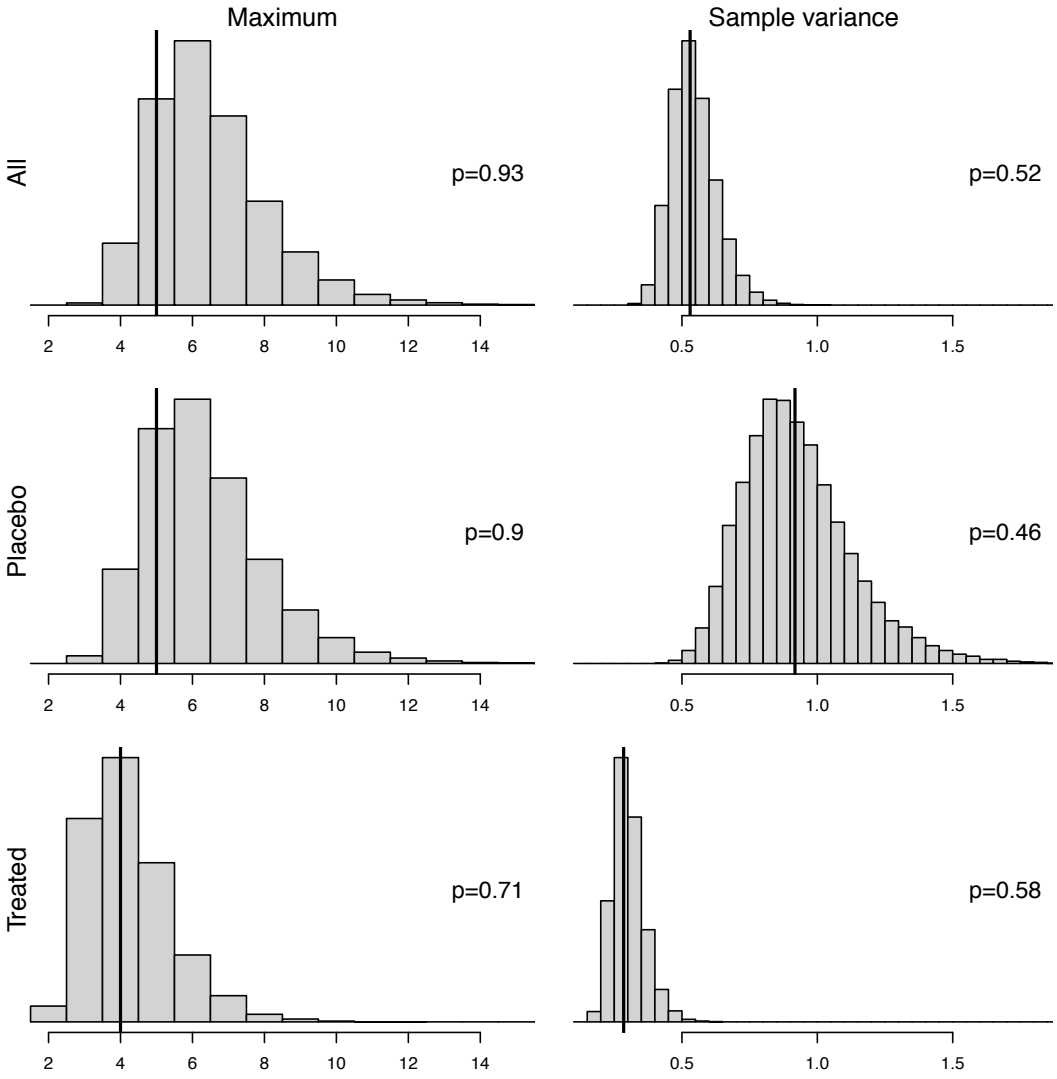

Figure A.2: Posterior predictive distribution of the maximum and sample variance for all subjects, subjects assigned to placebo, and subjects assigned to the active treatment. Solid vertical lines indicate the observed values of the test statistic.

in deviations from the 45 degree line. Neither graphical summary shows posterior predictive distributions that differ from the data distribution. We conclude that the model reflects the observed outcomes for each combination of treatment and observed indicator of rescue medication.

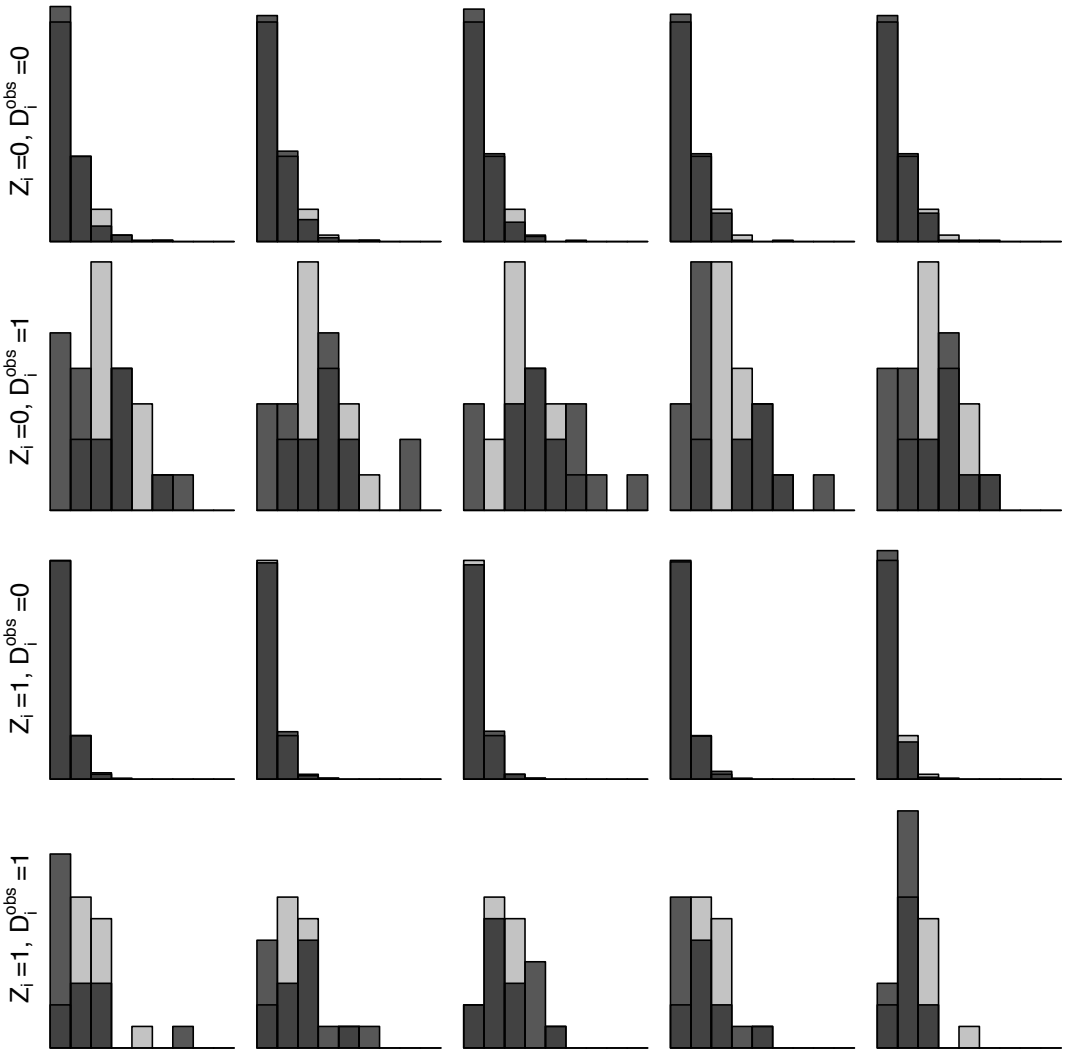

Figure A.3: Histograms of the posterior predictive distribution of observed outcomes overlaid against the histograms of the data distribution of observed outcomes. Posterior distributions are plotted in dark grey and the data distribution is plotted in light grey (note that overlap creates an even darker grey). Each column corresponds to an independent posterior predictive sample and each row corresponds to the labelled combination of treatment assignment and observed indicator of rescue medication. All x-axes range from zero to eight; y-axes are constant within row, but not across rows.

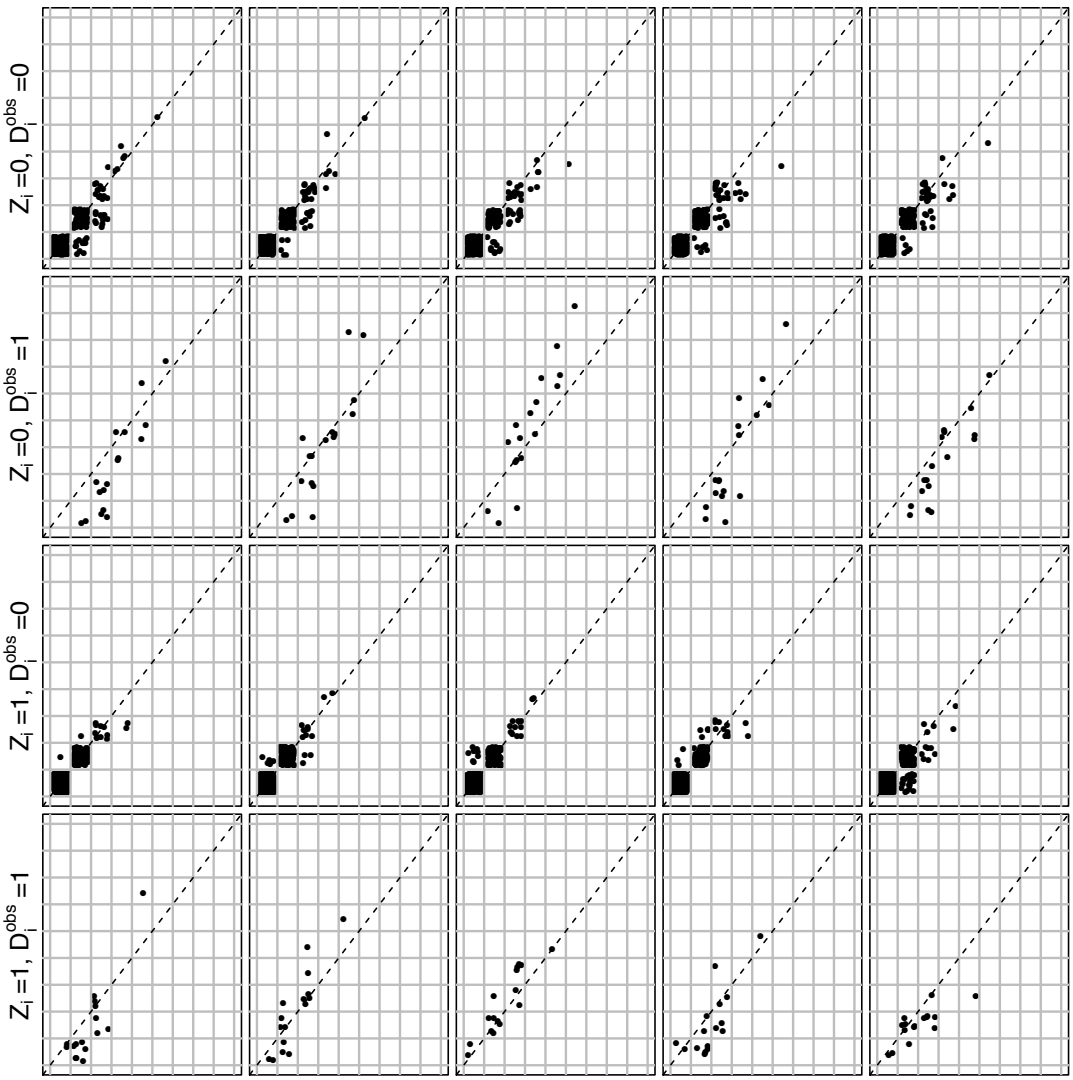

Figure A.4: QQ-plots of the data distribution of observed outcome (x-axis) and the posterior predictive distribution of observed outcomes (y-axis). Points are jittered to handle the several ties that occur in the discrete data. Each column corresponds to an independent posterior predictive sample and each row corresponds to the labelled combination of treatment assignment and observed indicator of rescue medication. Both x-axes and y-axes range from zero to eight with the grid delineating the discrete values.

Appendix B

Supplement to Chapter 2

We want to prove Equation (2.8) that states, given a test statistic of the aggregated observed data, $\mathbf{y}$, the posterior predictive p-value that conditions on the the entire observed data, $\mathbf{y}_{\text{obs}}$ and $\mathbf{W}$, is the same as the posterior predictive p-value that conditions on the aggregated data under the conditions that the prior distribution on the array of potential outcomes is placed on the sufficiently sharp partition Υ_0 and every element within each set of the partition treated symmetrically. The conditions state that

$$\mathrm{p}\big(\mathbf{Y}\big) = \begin{cases} \frac{\Pr\left(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\right)}{|\mathcal{Y}_0(N_*^{11})|} & \text{if } \mathbf{Y} \in \mathcal{Y}_0(N_*^{11}), \\ 0 & \text{otherwise,} \end{cases}$$

where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. If we show

$$\mathrm{p}\big(\mathbf{y}^{\text{pp}} \mid \mathbf{y}_{\text{obs}}, \mathbf{W}\big) = \mathrm{p}\big(\mathbf{y}^{\text{pp}} \mid \mathbf{y}\big), \tag{B.1}$$

the equality follows trivially.

We first apply the law of total probability to the left hand side of Equation (B.1):

$$\begin{aligned} \mathrm{p}\big(\mathbf{y}^{\mathrm{pp}} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big) &= \sum_{N_*^{11}} \left[\sum_{\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})} \mathrm{p}\big(\mathbf{y}^{\mathrm{pp}} \mid \mathbf{Y}\big)\ \mathrm{p}\big(\mathbf{Y} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big) \right] \\ &= \sum_{N_*^{11}} \mathrm{p}\big(\mathbf{y}^{\mathrm{pp}} \mid \mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big) \left[\sum_{\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})} \mathrm{p}\big(\mathbf{Y} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{W}\big) \right] \end{aligned} \tag{B.2}$$

We drop the term $\mathbf{y}_{\mathrm{obs}}$ and $\mathbf{W}$ from the first probability in the summation because given the array of potential outcomes, the predictive data will be independent of the observed data. Line (B.2) follows from Υ_0 being a sufficiently sharp partition for the aggregated observed outcomes. Showing the summation inside the brackets in line (B.2) equals

$$\Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11}) \mid \mathbf{y}\big)$$

will complete the proof. To see this fact, we argue that these two quantities are

proportional as follows:

$$\sum_{\mathbf{Y}\in\mathcal{Y}_0(N_*^{11})} \mathrm{p}\big(\mathbf{Y} \mid \mathbf{y}_{\text{obs}}, \mathbf{W}\big)$$

$$\propto \sum_{\mathbf{Y}\in\mathcal{Y}_0(N_*^{11})} \mathrm{p}\big(\mathbf{y}_{\text{obs}}, \mathbf{W} \mid \mathbf{Y}\big)\ \mathrm{p}\big(\mathbf{Y}\big) \tag{B.3}$$

$$= \frac{\Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big)}{\binom{N}{M}|\mathcal{Y}_0(N_*^{11})|} \sum_{\mathbf{Y}\in\mathcal{Y}_0^{\text{obs}}(N_*^{11})} 1 \tag{B.4}$$

$$= \frac{\Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big)}{\binom{N}{M}|\mathcal{Y}_0(N_*^{11})|} \sum_{\mathbf{M}_{0*}\in\mathcal{M}_{\mathbf{y}|\mathbf{N}_{0*}}} \left[\sum_{\mathbf{Y}\in\mathcal{Y}_0^{\text{obs}}(N_*^{11}):\mathbf{M}=\mathbf{M}_{0*}} 1\right] \tag{B.5}$$

$$= \frac{\Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big)}{\binom{N}{M}|\mathcal{Y}_0(N_*^{11})|} \sum_{\mathbf{M}_{0*}\in\mathcal{M}_{\mathbf{y}|\mathbf{N}_{0*}}} \binom{y_{\text{t}}}{M_{0*}^{11}}\binom{M-y_{\text{t}}}{M_{0*}^{00}}\binom{y_{\text{c}}}{N_*^{11}-M_{0*}^{11}}\binom{N-M-y_{\text{c}}}{N_{0*}^{00}-M_{0*}^{00}} \tag{B.6}$$

$$= \frac{\Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big)}{\binom{N}{y_{\text{t}},M-y_{\text{t}},y_{\text{c}},N-M-y_{\text{c}}}} \sum_{\mathbf{M}_{0*}\in\mathcal{M}_{\mathbf{y}|\mathbf{N}_{0*}}} \frac{\binom{N_{0*}^{00}}{M_{0*}^{00}}\binom{N_0^{01}}{M_{0*}^{01}}\binom{N_0^{10}}{M_{0*}^{10}}\binom{N_*^{11}}{M_{0*}^{11}}}{\binom{N}{M}} \tag{B.7}$$

$$\propto \Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big)\ \mathrm{p}\big(\mathbf{y} \mid \mathbf{Y} \in \mathcal{Y}_0(N_*^{11})\big) \tag{B.8}$$

$$\propto \Pr\big(\mathbf{Y} \in \mathcal{Y}_0(N_*^{11}) \mid \mathbf{y}\big) \tag{B.9}$$

where $\mathcal{Y}_0^{\text{obs}}(N_*^{11}) = \{\mathbf{Y} \in \mathcal{Y}_0(N_*^{11}) : \mathbf{y}_{\text{obs}} \text{ follows from } \mathbf{W} \text{ and } \mathbf{Y}\}$ and all proportionality statements imply a multiplicative constant that is a function of $\mathbf{y}_{\text{obs}}$ and $\mathbf{W}$. Line (B.3) follows by Bayes rule. Line (B.4) follows from the likelihood as defined in Equation (2.7) and the conditions on the prior distribution. Line (B.5) partitions the set $\mathcal{Y}_0^{\text{obs}}(N_*^{11})$ into the specific values of the treated unit aggregated potential outcomes, $\mathbf{M}$, which is a many-to-one function of $\mathbf{Y}$ and $\mathbf{Z}$. Line (B.6) counts the number of arrays of potential outcomes that result in $\mathbf{M}_{0*} = (M_{0*}^{00}, M_{0*}^{01}, M_{0*}^{10}, M_{0*}^{11})$ in combination with $\mathbf{Z}$. First consider the M treated units. There are y_{t} treated units

that are diseased that belong to either the doomed or harmed principal strata; there are

$$\binom{y_{\mathrm{t}}}{M_{0*}^{11}}$$

possible configurations for these y_{t} units. Similarly, the $M - y_{\mathrm{t}}$ treated units that are non-diseased belong to either the protected or immune principal strata; there are

$$\binom{M - y_{\mathrm{t}}}{M_{0*}^{00}}$$

possible configurations for these $M - y_{\mathrm{t}}$ units. The same argument for the y_{c} and $N - M - y_{\mathrm{c}}$ units in the control group provides the remaining terms in line (B.6). Line (B.7) rearranges terms and notes that

$$|\mathcal{Y}_0(N_*^{11})| = \binom{N}{N_{0*}^{00}, N_0^{01}, N_0^{10}, N_*^{11}}.$$

Line (B.8) recognizes the probability mass function of $\mathbf{y}$ in Equation (2.6). Line (B.9) follows from Bayes rule and completes the proof.

Appendix C

Supplement to Chapter 3

The simulated network of sexual contacts is broken into two elements. First, simulate the "relationships" in the network, that is the edges between nodes. Second, given a "relationship" exists, simulate the number of contacts for each relationship and the time at which each sexual contact occurred.

The network edges were sampled using the `degree.sequence.game` function in the `igraph` library in `R` (Csardi and Nepusz, 2006). The degree sequence was sampled from a power law distribution. Specifically, the probability of degree d was proportional to $d^{-2.5}$ for $d \in \{1, \ldots, 20\}$. The parametric model for the degree distribution was chosen to create a network similar to the observed network of sexual relations in Helleringer and Kohler (2007), in particular the proportion of small clusters and the number of units belonging to the smaller clusters and largest cluster.

Once the relationships were sampled, we next simulate the number of contacts and when they occurred. First, we simulate the start and end time of each relationship. The duration of the relationship was first sampled from an exponential random vari-

able with mean inversely proportional to the maximum number of degrees of the two units. For example, if two units have a relationship, and one partner has one degree and the other has two degrees, then the length of the relationship is a exponential random variable with mean $\mu/2$ where $\mu = 3$ corresponding to 3 years (the length of the study). The intuition is that units with more sexual partners will have shorter relationships. The average length of time for a monogamous relationship was 3 years. The midpoint of each relationship is randomly sampled to get the starting time and end time of each relationship, which are truncated to 0 and 3 respectively if they fall outside of the duration of the study. Every relationship now has a starting point and end point.

Given the start and end time of the each relationship, a Poisson random variable is drawn with exposure the length of time of the relationship (i.e. the end time minus the start time). For example, if a relationship started at time 0.5 and ended at time 1.1, the number of contacts is a Poisson random variable with mean 0.6λ where $\lambda = 60$ corresponds to the rate of sexual contacts per year in a relationship. If the number of contacts is greater than zero, the times of sexual contacts are uniformly drawn over the duration of the relationship. If the number of contacts is zero, we arbitrarily assign one contact occurring at the beginning of the relationship. All random variables are drawn independent of each other, so relationships may overlap and some units will be in multiple relationships at the same time.

The choice of parametric models was arbitrarily chosen, but the parameters settings were chosen to give reasonable results. Thus, an average of five sex acts per month was based on the results from Karim et al. (2010). Decomposing the network

simulation into first contacts and then time of sexual contact allowed more control over features of the network of sexual contacts over current dynamic network models (for a survey of network models, see Goldenberg et al., 2010). The dynamics of sexual networks are not well understood and we acknowledge that the results for our simulated population will be sensitive to the above choices.

Bibliography

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Baeten, J. M., D. Donnell, P. Ndase, N. R. Mugo, J. D. Campbell, J. Wangisi, J. W. Tappero, E. A. Bukusi, C. R. Cohen, E. Katabira, et al. (2012). Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine 367*(5), 399–410.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association 98*(462), 299–323.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association 72*(358), 355–366.

Bayarri, M. and J. Berger (2000). P values for composite null models. *Journal of the American Statistical Association 95*(452), 1127–1142.

Buchbinder, S. P., D. V. Mehrotra, A. Duerr, D. W. Fitzgerald, R. Mogg, D. Li, P. B. Gilbert, J. R. Lama, M. Marmor, C. del Rio, et al. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet 372*(9653), 1881–1893.

Cangul, M., Y. Chretien, R. Gutman, and D. Rubin (2009). Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Statistics in medicine 28*(20), 2531–2551.

Casella, G. and R. Berger (2001). *Statistical Inference*. Duxbury Press.

Cassell, M. M., D. T. Halperin, J. D. Shelton, and D. Stanton (2006). HIV and risk behaviour: Risk compensation: the Achilles' heel of innovations in HIV prevention? *British Medical Journal 332*(7541), 605.

Clark, C. E. (1961). The greatest of a finite set of random variables. *Operations Research 9*(2), 145–162.

Cohen, M. S. and L. R. Baden (2012). Preexposure prophylaxis for HIV—where do we go from here? *New England Journal of Medicine 367*(5), 459–461.

Cook, S. R., A. Gelman, and D. B. Rubin (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics 15*(3), 675–692.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.

Cox, D. R. and E. J. Snell (1989). *Analysis of binary data*, Volume 32. Chapman & Hall/CRC.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal, Complex Systems 1695*(5).

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics Letters B 195*(2), 216–222.

FDA (2001, May). *Guidance for Industry: E10. Choice of control group and related issues in clinical trials*. U.S. Department of Health and Human Services.

Fisher, R. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fox, R. J., D. H. Miller, J. T. Phillips, M. Hutchinson, E. Havrdova, M. Kita, M. Yang, K. Raghupathi, M. Novas, M. T. Sweetser, et al. (2012). Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *New England Journal of Medicine 367*(12), 1087–1097.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Freireich, E. J., E. Gehan, E. Frei, L. R. Schroeder, I. J. Wolman, R. Anbari, E. O. Burgert, S. D. Mills, D. Pinkel, O. S. Selawry, et al. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood 21*(6), 699–716.

Ganesh, A., L. Massoulié, and D. Towsley (2005). The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, Volume 2, pp. 1455–1466. IEEE.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika 52*(1-2), 203–223.

Gelman, A. (2003). A bayesian formulation of exploratory data analysis and goodness-of-fit testing*. *International Statistical Review 71*(2), 369–382.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A., X. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6*, 733–759.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.

Gilbert, P. B., R. J. Bosch, and M. G. Hudgens (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics 59*(3), 531–541.

Gold, R., G. Giovannoni, K. Selmaj, E. Havrdova, X. Montalban, E.-W. Radue, D. Stefoski, R. Robinson, K. Riester, J. Rana, et al. (2013). Daclizumab high-yield process in relapsing-remitting multiple sclerosis (SELECT): a randomised, double-blind, placebo-controlled trial. *The Lancet*.

Gold, R., L. Kappos, D. L. Arnold, A. Bar-Or, G. Giovannoni, K. Selmaj, C. Tornatore, M. T. Sweetser, M. Yang, S. I. Sheikh, et al. (2012). Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *New England Journal of Medicine 367*(12), 1098–1107.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning 2*(2), 129–233.

Grant, R. M., J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, P. Goicochea, M. Casapía, J. V. Guanira-Carranza, M. E. Ramirez-Cardich, et al. (2010). Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *New England Journal of Medicine 363*(27), 2587–2599.

Greenwood, M. and G. U. Yule (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine 8*(Sect Epidemiol State Med), 113.

Guest, G., D. Shattuck, L. Johnson, B. Akumatey, E. E. K. Clarke, P.-L. Chen, and K. M. MacQueen (2008). Changes in sexual risk behavior among participants in a PrEP HIV prevention trial. *Sexually Transmitted Diseases 35*(12), 1002–1008.

Halloran, M. E., I. M. Longini, M. J. Haber, C. J. Struchiner, and R. C. Brunet (1994). Exposure efficacy and change in contact rates in evaluating prophylactic HIV vaccines in the field. *Statistics in Medicine 13*(4), 357–377.

Halloran, M. E., I. M. Longini, and C. J. Struchiner (2010). *Design and analysis of vaccine studies*. Springer.

Halloran, M. E. and C. J. Struchiner (1995). Causal inference in infectious diseases. *Epidemiology 6*(2), 142–151.

Hammer, S. M., M. E. Sobieszczyk, H. Janes, S. T. Karuna, M. J. Mulligan, D. Grove, B. A. Koblin, S. P. Buchbinder, M. C. Keefer, G. D. Tomaras, et al. (2013). Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *New England Journal of Medicine 369*(22), 2083–2092.

Helleringer, S. and H.-P. Kohler (2007). Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. *Aids 21*(17), 2323–2332.

Hill, A. B. (1994). The continuing unethical use of placebo controls. *New England Journal of Medicine 331*, 394–398.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics 1*(1), 69–88.

Hoffman, M. D. and A. Gelman (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv 1111*(4246).

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Hudgens, M. G. and M. E. Halloran (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association 101*(473).

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imbens and Rubin (2014). *Causal Inference in Statistics and Social Sciences*. Unpublished.

Jin, H. and D. B. Rubin (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association 103*(481), 101–111.

Jones, J. H. and M. S. Handcock (2003). An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society of London. Series B: Biological Sciences 270*(1520), 1123–1128.

Kamb, M. L., M. Fishbein, J. M. Douglas Jr, F. Rhodes, J. Rogers, G. Bolan, J. Zenilman, T. Hoxworth, C. K. Malotte, M. Iatesta, et al. (1998). Efficacy of risk-reduction counseling to prevent human immunodeficiency virus and sexually transmitted diseases. *Journal of the American Medical Association 280*(13), 1161–1167.

Karim, Q. A., S. S. A. Karim, J. A. Frohlich, A. C. Grobler, C. Baxter, L. E. Mansoor, A. B. Kharsany, S. Sibeko, K. P. Mlisana, Z. Omar, et al. (2010). Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science 329*(5996), 1168–1174.

Kenah, E. and J. M. Robins (2007). Second look at the spread of epidemics on networks. *Physical Review E 76*(3), 036113.

Lasagna, L. (1979). Placebos and controlled trials under attack. *European Journal of Clinical Pharmacology 15*(6), 373–374.

Longini, I. M., J. S. Koopman, M. Haber, and G. A. Cotsonis (1988). Statistical inference for infectious diseases risk-specific household and community transmission parameters. *American Journal of Epidemiology 128*(4), 845–859.

Lublin, F. D. and S. C. Reingold (2001). Placebo-controlled clinical trials in multiple sclerosis: Ethical considerations. *Annals of neurology 49*(5), 677–681.

Meng, X. (1994). Posterior predictive p-values. *The Annals of Statistics*, 1142–1160.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical Review E 66*(1), 016128.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*(4), 558–625.

Neyman, J., D. Dabrowska, and T. Speed (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Nolen, T. and M. Hudgens (2011). Randomization-based inference within principal strata. *Journal of the American Statistical Association 106*(494), 581–593.

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549–556.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pitman, E. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society 4*(1), 119–130.

Pitman, E. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society 4*(2), 225–232.

Pitman, E. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika 29*(3/4), 322–335.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*(2), 191–199.

Polman, C., S. Reingold, F. Barkhof, P. Calabresi, M. Clanet, J. Cohen, G. Cutter, M. Freedman, L. Kappos, F. Lublin, et al. (2008). Ethics of placebo-controlled clinical trials in multiple sclerosis: a reassessment. *Neurology 70*(13 Part 2), 1134–1140.

Polman, C. H., P. W. O'Connor, E. Havrdova, M. Hutchinson, L. Kappos, D. H. Miller, J. T. Phillips, F. D. Lublin, G. Giovannoni, A. Wajgt, et al. (2006). A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *New England Journal of Medicine 354*(9), 899–910.

Rhodes, P. H., M. E. Halloran, and I. M. Longini Jr (1996). Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(4), 751–762.

Robins, J., A. van der Vaart, and V. Ventura (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association 95*(452), 1143–1156.

Rosenbaum, P. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika 88*(1), 219–231.

Rosenbaum, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association 91*(434), 465–468.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rubin, D. (1978a). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.

Rubin, D. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. (1978b). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the Section on Survey Research Methods*, pp. 20. The Association.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with censoring due to death. *Statistical Science 21*(3), 299–309.

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys*, Volume 307. Wiley.

Rubin, D. B. et al. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*(4), 1151–1172.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Sommer, A. and S. Zeger (1991). On estimating efficacy from clinical trials. *Statistics in Medicine 10*(1), 45–52.

Stan Development Team (2013). Stan: A c++ library for probability and sampling, version 2.0.

Temple, R. and S. S. Ellenberg (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine 133*(6), 455–463.

Temple, R. J. (1994). Special study designs: early escape, enrichment, studies in non-responders. *Communications in Statistics-Theory and Methods 23*(2), 499–531.

Tenser, R. B. (2009). Ethics of placebo-controlled clinical trials in multiple sclerosis: a reassessment. *Neurology 72*(13), 1191–1192.

Thigpen, M. C., P. M. Kebaabetswe, L. A. Paxton, D. K. Smith, C. E. Rose, T. M. Segolodi, F. L. Henderson, S. R. Pathak, F. A. Soud, K. L. Chillag, et al. (2012). Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. *New England Journal of Medicine 367*(5), 423–434.

Van Damme, L., A. Corneli, K. Ahmed, K. Agot, J. Lombaard, S. Kapiga, M. Malahleha, F. Owino, R. Manongi, J. Onyango, et al. (2012). Preexposure prophylaxis for HIV infection among African women. *New England Journal of Medicine 367*(5), 411–422.

Volz, E. and L. A. Meyers (2009). Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Interface 6*(32), 233–241.

Walker, B. D. and D. R. Burton (2008). Toward an AIDS vaccine. *Science 320*(5877), 760–764.

Welch, B. (1937). On the z-test in randomized blocks and latin squares. *Biometrika 29*(1/2), 21–52.

White, I. R., C. Bamias, P. Hardy, S. Pocock, and J. Warner (2001). Randomized clinical trials with added rescue medication: some approaches to their analysis and interpretation. *Statistics in Medicine 20*(20), 2995–3008.

White, I. R., J. Carpenter, S. J. Pocock, and R. A. Henderson (2003). Adjusting treatment comparisons to account for non-randomized interventions: an example from an angina trial. *Statistics in Medicine 22*(5), 781–793.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics 28*(4), 353–368.